
THE UNIVERSITY
*of* EDINBURGH

# Development of computational methods to analyse single-cell RNA-sequencing data of γδ-T cells in human peripheral blood and breast cancer

Katerina Boufea


THE UNIVERSITY
*of* EDINBURGH

Thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
to the University of Edinburgh
2020

# Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Katerina Boufea

April 2020

# Abstract

With the improvement of sequencing protocols and decreasing sequencing costs, single-cell RNA sequencing has become widely accessible. Cells, once thought to be of the same type based on their location or morphology, are increasingly found to be heterogeneous with respect to gene expression levels. Computational analysis of single cell transcriptome data allows the identification of novel and rare cell populations and cell states, as well as the comparison of cell populations across tissues and conditions. However, resolving cell types and comparison of scRNA-seq data across datasets are challenging due to technical factors such as sparsity, low numbers of cells and batch effects. To address these challenges, I developed scID, which uses the Fisher's Linear Discriminant Analysis-like framework to identify transcriptionally related cell types between scRNA-seq datasets. I demonstrate the accuracy and performance of scID relative to existing methods on several published datasets. By increasing power to identify transcriptionally similar cell types across datasets showing batch effects, scID enhances an investigator's ability to integrate and reveal development-, disease- and perturbation-associated changes in scRNA-seq data. Using scID and other methods for data alignment, unsupervised clustering and differential gene expression analysis, I explored the heterogeneity within γδ-T cells from human peripheral blood and breast tumour samples from three healthy donors and two breast cancer patients. Two δ1 and three δ2 subtypes of γδ-T cells were identified in blood and one δ1 and two δ2 subtypes of γδ-T cells in breast

tumour. These subtypes differed in antigen presentation, cytotoxicity, and IL17 and IFNγ production. Compared to blood γδ-T cells, breast tumour-infiltrating γδ-T cells were more activated and expressed higher levels of cytotoxic genes, yet were immunosuppressed. A breast tumour subtype that was δ1 and IFNγ positive had no obvious similarity to any subtype observed in blood γδ-T cells and was the only subtype associated with improved overall survival of breast cancer patients. An additional method for overcoming batch effects and enabling comparison of cell populations across donors and conditions is to pool cells from multiple donors within a single scRNA-seq experiment. Experimental methods that enable the tracking of donor identity of each cell require heavy manual processing and are costly. Computational methods, on the other hand, can determine donor identities of cells based on genetic variants. Technical factors, such as sparsity and gene fragment capture, as well as biological factors, such as cell-type-specific gene expression, can present challenges. In the last chapter I explored the use of deep learning to implement a Non-negative matrix factorization method that clusters cells based on genetic variants and identifies donor-specific genetic variants that can be used for validation.

# Lay Summary

A single cell is the smallest and fundamental unit of life. Even though all human cells carry the same DNA, they differ in morphology and function between tissues and conditions. A new experimental approach, called single-cell RNA-sequencing, enables researchers to measure the expressed genes in individual cells. This information can be used to group cells based on common gene expression patterns and define different cell types in complex tissues. Cancer research can take advantage of such information to study how our immune cells function in the presence of tumour cells. To understand how our immune system is affected by the presence of cancer cells, we need to compare equivalent cell populations across healthy and diseased samples. During my PhD I worked on the development of computational methods to enable such analysis. I developed a method, called scID, to identify cells that express the same sets of genes across different datasets. I then used scID and other published methods to identify subtypes of an immune cell type called γδ-T cells in blood from healthy donors and breast tumour samples. This showed five different subtypes of γδ-T cells in blood and three different subtypes of γδ-T cells in breast tumour. Two of the subtypes were common between the two conditions while the other breast tumour γδ-T cell subtype was not observed in blood samples. Interestingly this subtype was associated with better survival rates of patients with breast tumour. An alternative way of comparing cells between different conditions or individuals is to pool cells from multiple donors together before the experiment. Then we

need to know the donor identity of each cell in order to ask questions such as whether all donors have the same cell types or whether there is a difference in the proportion of each cell type between the donors. However, tracking the donor identity of the cells in the single cell RNA-sequencing experiment is inefficient and expensive. Here I present a method to computationally label the cells. To achieve this I extract information on genetic variants of each cell. These are positions on the genome whether the genomic sequence of a cell differs from a reference genome. Some of them are donor-specific, thus cells that have the same variants are expected to be from the same donor.

# Acknowledgements

First and foremost, I would like to sincerely thank my supervisor Dr. Nizar Batada for giving me this opportunity to join his lab and for guiding me with kindness and patience throughout this scientific journey. I owe him a huge debt of gratitude for teaching me most of what I know now, pointing me to directions where I can have an actual impact with my research but also giving me the freedom to adjust the project to my skills and interests. Without him, this work would not be the same.

Similarly, I express my gratitude to my co-supervisor Prof. Chris Ponting for sparing his valuable time for fruitful meetings and suggestions. His insightful comments and support during the writing process incented me to widen my research from various perspectives and contributed to the improvement of the clarity of this thesis.

Besides my supervisors, I would like to thank the rest of my thesis committee and especially Prof. Chris Haley for assisting me with my annual reviews and for helping me overcome other difficulties that presented themselves during this time.

I would also like to thank Dr. Sohan Seth from the School of Informatics for his valuable comments and suggestions in the scID method, Dr. Victor González-Huici from the Batada lab for performing all single-cell RNA-sequnecing of the

γδ-T cell samples and Marcus Lindberg for the FACS validation of the γδ-T cell subtype biomarkers.

This journey was shared with many old and new friends, within the institute and outside. I deeply thank all of them for helping me or stimulating me as we were going through the same process or for just making me enjoy my time and being part of my life.

Undoubtedly, I am forever indebted to my parents who have provided me with every opportunity to succeed by encouraging and supporting me throughout all these years of my studies. They have always given me freedom to choose what I want to do in my life and without them I would never be where I am now. Finally, I would like to thank my partner for his love, encouragement and understanding through this process, and his endless patience with me. I would have not been here without him.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**ARI** Adjusted Rand Index

**BRCA** Breast Cancer

**cDNA** complementary DNA

**CPM** Counts Per Million mapped reads

**DNA** Deoxyribonucleic Acid

**ERCC** External RNA Controls Consortium

**FACS** Fluorescence-Activated Cell Sorting

**FPR** False Positive Rate

**GO** Gene Ontology

**HVG** Highly Variable Gene

**ICA** Independent Component Analysis

**KNN** K-Nearest Neighbours

**LDA** Linear Discriminant Analysis

**MHC** Major Histocompatibility Complex

**mRNA** messenger RNA

**NB** Negative Binomial

**NMF** Non-negative Matrix Factorisation

**PBMC** Peripheral Blood Mononuclear Cell

**PCA** Principal Component Analysis

**PCR** Polymerase Chain Reaction

**PDF** Probability Density Function

**qPCR** Quantitative Polymerase Chain Reaction

**RNA** Ribonucleic Acid

**RPKM** Reads Per Kilobase of transcript per Million mapped reads

**SNN** Shared Nearest Neighbours

**SNP** Single Nucleotide Polymorphism

**TCGA** The Cancer Genome Atlas

**TCR** T Cell Receptor

**TNBC** Triple Negative Breast Cancer

**TPM** Transcripts Per Million mapped reads

**TPR** True Positive Rate

**tSNE** t-distributed Stochastic Neighbour Embedding

**UMAP** Uniform Manifold Approximation and Projection

**UMI** Unique Molecular Identifier

**UTR** Untranslated Region

**VAE** Variational Autoencoder

**VI** Variation of Information

**ZINB** Zero-Inflated Negative Binomial

# Chapter 1

# Introduction

A single cell is the smallest and fundamental unit of life, often called the "building block of life". In complex multicellular organisms, such as the human, cells are specialised into different cell types with distinct morphological and functional characteristics. Understanding the different functions of our cells is key to studying human biology. However, all cells essentially carry the same Deoxyribonucleic Acid (DNA). So what causes this heterogeneity between tissues and cells?

Initial attempts to identify distinct cell types were focusing on defining them based on their location, morphology and interactions with other cells, as observed under the microscope. The development of immunohistochemistry (Coons et al., 1941) and Fluorescence-Activated Cell Sorting (FACS) (Fulwyler, 1965) enabled measurement of specific cell surface proteins and revealed transcriptional differences between cells with the same morphology. Although the human genome has 20,000 to 25,000 genes (Collins et al., 2004), only a small fraction of them are expressed in each cell. Genes and their variable expression levels can explain the observed heterogeneity between cell types and tissues.

Ribonucleic Acid (RNA)-sequencing enabled researchers to study the differences in gene expression levels that explain the specialised functions of different cell types. However, RNA is a much more unstable molecule compared to DNA, thus RNA-sequencing became possible only after the development of reverse transcription polymerase chain reaction (RT-PCR) that enabled the synthesis of complementary DNA (cDNA) from RNA making PCR analysis of RNA molecules possible (Freeman et al., 1999). Another limitation of RNA-sequencing compared to DNA-sequencing, is the low amount of input material to the sequencer. For example, a human cell contains less than 1 pg of mRNA (Kawasaki, 2004), while most RNA sequencing kits require between 10 and 400 ng of input mRNA material. Thus, RNA-sequencing was first used to measure gene expression levels in samples consisting of a collection of cells, known as bulk RNA-sequencing. Bulk RNA-seq data can be used to identify differentially expressed genes between different samples or cell types by comparing the average gene expression levels from a homogeneous set of cells. This leads to loss of information on the intra-sample variance of gene expression, assuming all cells have similar transcriptional patterns. Additionally, this requires availability of known markers for isolation of cells of a specific type, thus it does not allow identification of novel subpopulations.

Advances in DNA sequencing and the development of single-cell Quantitative Polymerase Chain Reaction (qPCR) (Bengtsson et al., 2008), have enabled the measurement of gene expression at the single cell level, allowing for the characterisation of genetic variability of heterogeneous samples. Single-cell RNA-sequencing can unravel the heterogeneity of cell populations, especially rare ones, that was masked until now in bulk RNA-sequencing studies. Discovery of previously unknown cell types can be unbiased without requiring prior knowledge of markers. Gene expression differences between populations can now be identified based on comparisons of the distributions of gene expression between samples, allowing for better identification of biomarkers and functional

annotation. Additionally, single-cell RNA-sequencing allows identification of transitioning states of cell types during development.

The most promising application of single-cell RNA-sequencing is The Human Cell Atlas (Regev et al., 2017). Several labs around the world are collaborating by generating good quality single-cell RNA-sequencing data from different tissues and providing annotated cell types along with their gene expression profiles. By combining such information with existing location and morphological information of the different cell types this can serve as a valuable public resource for the whole scientific community for improving our understanding of the heterogeneity of human cells and tissues in health and disease. A similar effort to map the heterogeneity of mouse tissues is the Mouse Cell Atlas (Han et al., 2018).

## 1.1 Experimental Protocols for single-cell RNA-sequencing

Since the first single-cell RNA-sequencing data were obtained in 2009 (Tang et al., 2009), several improvements in protocols have led to broadly accessible and robust methods. There are two main categories of protocols for single-cell RNA-sequencing, plate-based and bead-based. The main difference between these two lies in the cell capturing method. Plate-based methods, such as CEL-Seq (Hashimshony et al., 2012), Smart-Seq2 (Picelli et al. (2013), Picelli et al. (2014)), MARS-Seq (Jaitin et al., 2014) and CEL-Seq2 (Hashimshony et al., 2016), require manual pipetting of the cells in multi-well plates. On the other hand, with bead-based technologies thousands of cells can be processed in one experiment by encapsulating them into droplets (Klein et al. (2015), Macosko et al. (2015), Zheng et al. (2017)) or into wells (Han et al. (2018), Gierahn et al. (2017)) with barcoded beads. The best choice of protocol depends on the research questions investigated.

Bead-based methods can yield thousands of cells as they require less manual handling, whereas plate-based methods are restricted to hundreds of cells sorted into 96- or 384-well plates. The difference in the number of cells processed leads to a significant difference in the library sizes, with plate-based methods detecting more genes per cell and allowing the identification of more subtle differences between cell subpopulations compared to bead-based. Thus, for uncovering unknown heterogeneity in cell populations or studying rare populations bead-based methods can be more suitable resulting in a sufficient number of cells to improve accuracy of unsupervised clustering (Kiselev et al., 2019). For identifying differentially expressed genes and subtle differences between cell populations, plate-based methods can be more appropriate by capturing more genes expressed in each cell.

**Table 1.1** summarizes the main characteristics of all experimental protocols used to generate the datasets presented in this thesis. These protocols will be further discussed in the following sections.

**Table 1.1:** Summary of main features of single-cell RNA-sequencing experimental protocols. UMI, Unique Molecular Identifier; UTR, Untranslated Region

| Protocol | Publication | Platform | Genomic region | UMI |
|---|---|---|---|---|
| **Smart-Seq2** | Picelli et al. (2013) | Plate-based | full-length | No |
| **CEL-Seq** | Hashimshony et al. (2012) | Plate-based | 3' UTR | Yes |
| **CEL-Seq2** | Hashimshony et al. (2016) | Plate-based | 3' UTR | Yes |
| **Chromium 10X** | Zheng et al. (2017) | Bead-based | 3' UTR | Yes |
| **Drop-seq** | Macosko et al. (2015) | Bead-based | 3' UTR | Yes |

### 1.1.1 Plate-based methods

In this thesis I will be using data generated with three different plate-based protocols: Smart-Seq2 (Picelli et al., 2013), CEL-Seq (Hashimshony et al., 2012) and CEL-Seq2 (Hashimshony et al., 2016). Smart-Seq2 is a low-throughput method that requires little special equipment but intense manual work. Cells are placed into 96- or 384-well plates and lysed. To increase the amount of RNA, the extracted RNA is reverse transcribed to complementary DNA (cDNA) and cDNA is then amplified by Polymerase Chain Reaction (PCR). The amplified cDNA of each cell/well is fragmented and adaptors are added to each fragment to prepare the libraries for sequencing. The main advantage of Smart-Seq2 is the ability to capture the full-length messenger RNA (mRNA), offering good coverage of the transcriptome, including rare transcripts, while all other methods mentioned here only capture the 3' Untranslated Region (UTR). Thus, Smart-Seq2 is a suitable approach to study gene isoforms and genetic variants.

The CELSeq (Hashimshony et al., 2012) flowchart is very similar to that of Smart-Seq2 but it additionally integrates the use of unique molecular identifiers (UMIs) into the primer to enable identification of PCR amplification bias. CEL-Seq2 (Hashimshony et al., 2016) is an improved protocol over CEL-Seq with increased sensitivity, decreased hands-on time and lower price.

### 1.1.2 Bead-based methods

Chromium 10X (Zheng et al., 2017) was used in our lab to generate the datasets in **Chapter 3**. In the 10X platform, cells mixed with barcoded beads and reagent are encapsulated in partitioning oil to form single nanoliter droplets, called Gel Beads in Emulsion. Cell capturing with Drop-seq (Macosko et al., 2015) is similar to the 10X workflow, however in 10X reverse transcription is carried out within the droplets, while in Drop-seq reverse transcription is carried out after demulsification. Although 10X has higher sensitivity than Drop-seq, the significantly lower cost of Dropseq makes it a preferred approach for processing very high numbers of cells (Zhang et al., 2019a).

## 1.2 Computational methods for the analysis of single-cell RNA-sequencing data

The most common analysis of single-cell RNA-sequencing data follows the flowchart of **Figure 1.1**. The first step involves sequence alignment to a reference genome and count of the expressed genes in each cell. This returns a raw counts gene expression matrix with genes in rows and cells in columns. Each value indicates the number of reads mapped to a specific gene in a specific cell. To correct for gene expression bias and differences in library depths between different cells, gene counts are normalised with different methods. This normalised data is used for clustering, which also involves feature extraction and dimensionality reduction, to identify transcriptionally distinct populations in the data. The identified clusters of cells are visualised in a two dimensional projection and differential expression analysis is used to identify cluster-specific genes.

### 1.2.1 Alignment and counts

Depending on the type of protocol used to generate the data, different alignment methods can be more appropriate. Commonly used methods are the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) and STAR (Dobin et al., 2013), although not specifically developed for single-cell RNA-sequencing data. Alternatively, faster but accurate quantification of gene expression can be obtained using `kallisto` (Bray et al., 2016) that performs pseudoalignment of reads identifying the transcript origin of each read without the specific genomic positions.

When UMIs are available, they can be used to collapse duplicated reads due to amplification bias into a single read. Throughout my PhD I have been using the CellRanger pipeline from 10X Genomics (Zheng et al., 2017) to demultiplex the

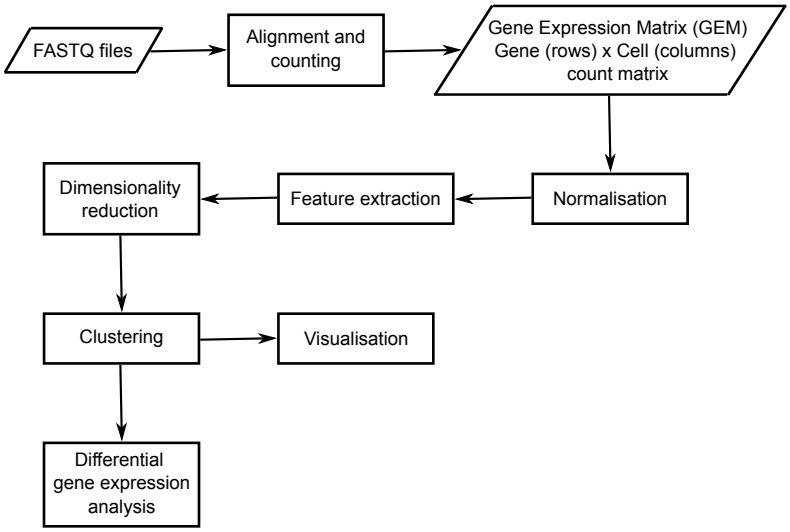



**Figure 1.1:** Flowchart of common analysis steps of single-cell RNA-sequencing data. The pipeline starts with the alignment of fastq files and counting of reads per cell corresponding to each gene, resulting in a counts gene expression matrix where rows represent genes and columns cells. Raw counts are then normalised to correct for technical differences between cells and allow comparison of gene expression across cells. The next aim is to cluster the data to identify groups of transcriptionally distinct cell populations, which also involves feature extraction and dimensionality reduction. Clusters of cells can then be visualised and finally, differentially expressed genes can be found for the identified clusters of cells.

raw base call files from Illumina sequencers to `fastq` files and align the reads and obtain the gene counts matrix. CellRanger uses STAR (Dobin et al., 2013) to align the data and additionally incorporates methods to filter low quality cells based on the number of reads and UMIs.

### 1.2.2 Multiplet detection

One of the disadvantages of high-throughput isolation and barcoding methods is the generation of multiplets which arise when two or more cells are co-captured and encapsulated in a single bead or well with the same cell barcode. These

libraries, called multiplets, represent hybrid gene expression profiles and can affect the computational analysis of the data as well as the interpretation of the results. Multiplets often form separate clusters which can be interpreted as an intermediate biological state between the actual states of the cells that comprise them. Additionally, they can affect the results of differential expression analysis by reducing the gene expression difference and statistical significance of truly differentially expressed genes. Some technical characteristics of multiplets are the high number of transcripts and the co-expression of markers of different cell types. The latter however, can not be applied to any project, since it requires availability and knowledge of highly expressed and exclusive markers for the cell types that comprise the multiplets. Filtering libraries based on the observed number of genes is also not a robust and systematic methods as it is highly dependent on the manual inspection of the distribution of genes per library in the data and is highly subjective.

Computational methods have been developed for the automatic identification of multiplets from single-cell RNA-sequencing data. Scrublet (Wolock et al., 2019) is focusing on the identification of doublets, i.e. libraries consisting of two cells, since these make up the majority of multiplets. Artificial doublets are calculated as linear combinations of random pairs of cells from the data and K-Nearest Neighbours (KNN) graphs are constructed to calculate the similarity of each cell to the simulated doublets, which is used for the classification of the cells as doublets or singlets, i.e. libraries that represent a single cell. A very similar approach is implemented in DoubletFinder (McGinnis et al., 2019). Artificial doublets are generated by averaging the gene expression profiles of random pairs of cells from the data and again doublets are selected based on their similarity to the artificial doublets.

All methods identify an expected number of doublets based on the experimental design. This however, can be overly stringent or too relaxed based on the quality

of each data set. It is thus advised that the labelled doublets are only removed from the data after clustering and differential expression analysis, if they are found to separate from the other cells or confound downstream analysis.

### 1.2.3 Normalisation

Normalisation of raw count data is essential to eliminate cell-specific bias and decrease false detection rates when comparing gene expression across different cells. The read counts of all genes in a cell are expected to be proportional to the expression levels of the genes and stochastic cell-specific factors. Two common approaches for normalisation are Counts Per Million mapped reads (CPM) and Transcripts Per Million mapped reads (TPM) mapped reads that have been developed for bulk RNA-sequencing data. CPM is used with edgeR for bulk RNA-sequencing data analysis where the read count $X_i$ of a gene $i$ in a cell is scaled by the total number of reads $N$ in the cell times one million, assuming all cells have equal molecules of mRNA and count depth differences between cells are due to sampling.

$$CPM_i = \frac{X_i}{N} 10^6 \tag{1.1}$$

While CPM is a between-samples normalisation method where gene expression is normalised by the library depth of each cell and allows comparison of gene expression across cells, TPM (Li and Durbin, 2009) is a within-sample normalisation allowing comparison of different genes' expression within a cell. TPM normalizes the gene counts by the gene's and transcript's length. The TPM-normalised expression of gene $i$ is given by:

$$TPM_i = \frac{X_i}{\tilde{l_i} \sum_j \frac{X_j}{\tilde{l_j}}} 10^6 \tag{1.2}$$

where $\tilde{l_i}$ is the effective length of gene $i$ given by:

$$\tilde{l_i} = l_i + \mu_{FLD} + 1 \tag{1.3}$$

where $\mu_{FLD}$ is the mean fragment length.

CPM and TPM are linear normalisation methods developed for bulk RNA-sequencing data. However, single-cell specific technical bias, such as zero-inflation and dropouts, are not accounted for in these approaches. Some methods proposed for the normalisation of single cell RNA-seq data apply cell-specific normalisation based on estimated size factors. Brennecke et al (Brennecke et al., 2013) use a cell-specific normalisation factor calculated as follows. The geometric mean of each gene across all cells is calculated and for each cell the median of the ratio of the cell's gene counts to these geometric means defines a cell-specific size factor. Finally, the gene expression counts of each cell are normalised by dividing the read counts by the cell-specific size factor. BASiCS (Vallejos et al., 2015) uses a Bayesian hierarchical model to normalize the data by estimating the cell size factors from External RNA Controls Consortium (ERCC) molecules, also known as spike-ins. Spike-ins are synthetic RNA molecules inserted in the cells and used to measure the detection rate of RNA-seq experiments (Jiang et al., 2011). Instead of spike-ins, Linnorm (Yip et al., 2017) utilizes a set of automatically detected stably expressed genes to estimate the technical bias between cells.

Alternative methods use probabilistic models to apply not only cell-specific but also gene-specific normalisation of read counts. scVI (Lopez et al., 2018) models the read counts of a gene as a sample drawn from a Zero-Inflated Negative Binomial (ZINB) distribution. Hafemeister and Satija (Hafemeister and Satija, 2019) on the other hand show that Negative Binomial (NB) and ZINB models lead to overfitting of single-cell RNA-seq data and suggest a regularised negative

binomial regression. The regularisation is achieved by pooling information (geometric mean) across genes with similar average expression. Finally, log transformation is applied to reduce skewness when downstream analysis methods assume normally distributed data.

Cole et al (Cole et al., 2019) show that different datasets require different normalisation methods and suggest a method called scone for identifying the most suitable normalisation method for a given single-cell RNA-seq dataset.

### 1.2.4 Imputation

Up to 80% of the expression values of the data can be zeros, which causes challenges in using standard normalisation methods and similarity metrics. The term "dropout" is used to describe the case where no reads corresponding to a gene are observed in a cell leading to a zero value. There are two cases that lead to dropouts: either a gene is not expressed in a cell (biological zero) or a gene is expressed in a cell but due to poor RNA capture efficiency or technical variation it is not captured and measured (technical zero). Methods have been developed to impute values for technical zeros.

One approach for imputation is to use generative models to determine whether an observed zero is biological or technical and impute values for technical zeros from the other cells in the data. Examples of these approaches are BISCUIT (Prabhakaran et al., 2016), CIDR (Lin et al., 2017) and scImpute (Li and Li, 2018). Other methods, such as MAGIC (van Dijk et al., 2017), apply a smoothing of gene expression for all genes (dropouts and non-dropouts) based on pooled gene expression from a set of "similar" cells. Finally, scImpute (Li and Li, 2018), EnImpute (Zhang et al., 2019b) and scVAE (Grønbech et al., 2019) are examples of methods that attempt to reconstruct the data using machine learning.

The main concern regarding imputation is the circularity of the problem that might lead to overimputation (Lähnemann et al., 2019). Imputation of gene expression values for biological zeros can lead to wrong clustering of the data and false positives in differential gene expression analysis. Thus, it is suggested that imputation is avoided when possible and account for sparsity with appropriate statistical models instead (Lähnemann et al., 2019).

### 1.2.5 Feature Extraction

According to the Human Genome Project, the human genome has 20,000 to 25,000 genes (Collins et al., 2004). However, only a small fraction of them are expressed in a cell, depending on its cell type. Of those expressed genes, only a small fraction of genes can distinguish the different cell types, since many genes that are involved in generic cell functions are shared between cells of different cell types (housekeeping genes). Thus, single-cell analysis methods that aim to identify and characterize the different cell populations within a single-cell RNA-seq dataset start by extracting those informative features in order to reduce the dimensionality of these highly sparse datasets.

The most common approach of feature extraction is selection of highly variable genes. Highly variable genes are selected based on the relationship between average expression and variance for analysis with Seurat (Hoffman et al., 2018). It is known that variance depends on average expression level of a gene, with variance increasing for lowly expressed genes (Grün et al., 2014). To account for this mean-variance relationship, Stuart et al. (2019) apply a variance-stabilizing transformation prior to fitting a linear regression between average expression and variance. Hao et al. (2019) suggest a similar approach but the genes are first categorised into bins based on their average expression levels and a bin-specific linear regression is fitted between the coefficient of variation and the mean gene expression. Other methods, such as M3Drop (Andrews and Hemberg, 2019) and

scmap (Kiselev et al., 2018), select features based on the relationship between average expression and dropout rate.

All the above methods select a user-defined number of top highly variable genes (HVGs). However this number is not intuitive and there is no systematic way to estimate a suitable number according to each dataset. Systematic evaluation of several methods of selection of highly variable genes has shown very small overlap between HVGs selected with different methods and unstable performance across different datasets as evaluated by the result of unsupervised clustering (Yip et al., 2018).

### 1.2.6 Dimensionality Reduction

Even though feature extraction can reduce the dimensions of the data for clustering or other downstream analysis, dimensionality reduction methods are additionally used to obtain a low-dimension representation of the data that can uncover the underlying biological characteristics and reduce the noise. It has been shown that fewer dimensions than the total number of genes measured in a single-cell RNA-seq dataset are sufficient to describe the transcriptional programs of the cell populations (Heimberg et al., 2016).

The reduced dimensions are linear or non-linear combinations of the original features (genes). Principal Component Analysis (PCA) is the most commonly used linear dimensionality reduction method (Butler et al. (2018), Kiselev et al. (2017)). Principal components are selected based on their contribution to the variation of the data, selecting those components with highest contribution. A heuristic method to identify the number of principal component required is the use of an "elbow" plot, where components are plotted against the percentage of variance they explain. The number of principal components is selected near the "elbow" of the curve, where the variance becomes stably low for any new

component included. Another method, called jackstraw, uses a permutation test. Permuted subsets of the data are used to run PCA and construct a null distribution of principal components scores and test whether a gene is significantly associated to a principal component as well as determine the number of sufficient components (Macosko et al. (2015), Chung and Storey (2015)).

### 1.2.7 Clustering

Clustering is a typical step in single-cell RNA-seq data analysis. Organizing the cells based on shared transcriptional patterns helps identify biologically and functionally distinct populations in a sample. There are several different clustering methods and tools available. SC3 (Kiselev et al., 2017) and SIMLR use $k$-means clustering to group the cells into $k$ clusters by minimizing the distance of each cell to the centroid of the cluster it is assigned to and optimizing the centroids to be distant from each other. SC3 (Kiselev et al., 2017) uses a combination of Euclidean, Pearson and Spearman distance metrics to measure the similarity between cells. SIMLR on the other hand uses Gaussian kernels based on the Euclidean distance between cells to construct the similarity matrix for clustering.

A graph-based approach is implemented in Seurat (Butler et al., 2018). Graph-based clustering, such as the KNN and the Shared Nearest Neighbours (SNN) graphs (Xu and Su, 2015), represent the data as graphs with cells being the nodes connected by weighted edges, with weights representing a measure of similarity between two cells. In Seurat, a graph is constructed based on Euclidean distance in PCA space and the weight between two cells is calculated based on the overlap of the sets of their neighbours (Jaccard distance). Clusters are identified in the constructed neighbourhood using the Louvain method for community detection (Blondel et al., 2008), where communities are detected so that the density of the network within a community is maximised compared to the density of a random network.

### 1.2.8 Visualisation

Visual inspection of the data can be achieved by projection of the cells onto two dimensions. Due to the high dimensionality of the data, dimensionality reduction is again required to identify the underlying manifold of the data in order to place transcriptionally similar cells closer together.

t-distributed Stochastic Neighbour Embedding (tSNE) has been shown to preserve the local structure of high-dimensional data data (Van Der Maaten and Hinton, 2008). The low-dimensional space resulting from tSNE represents pairwise similarities between cells, where the similarity between a cell $c_i$ and a cell $c_j$ represents the probability of $c_j$ being selected as a neighbour of $c_i$ if neighbours of $c_i$ were selected from a probability density function under a Student-t distribution centred at $c_i$.

A modification of t-SNE is the Uniform Manifold Approximation and Projection (UMAP) that computes a fuzzy topological representation of the original data (McInnes et al., 2018). The set of cells is represented as a graph network and cells (nodes) are connected to their nearest neighbours via edges with weights representing the similarity between them. Becht et al. (2019) claim that UMAP can better capture the global structure of the data compared to t-SNE. This means that while projection of cells with both t-SNE and UMAP shows information on the homogeneity of the cells within a cluster by their closeness in a two dimensional space, between-cluster distances are better reflected by UMAP compared to t-SNE. However, comparison of the two algorithms showed that they both preserve the global structure of the data equally well and both algorithms' performance is highly dependent on the initialisation parameters (Kobak and Linderman, 2019).

### 1.2.9 Differential gene expression analysis

Differential gene expression analysis can reveal cluster-specific changes in gene expression between different cell types or states across different conditions. This can help understand the role of specific cell types in development (Karaiskos et al., 2017), disease (Stubbington et al., 2017) and drug response (Kang et al. (2018), Kim et al. (2015)). Methods such as DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2009), and limma (Wu et al., 2015) developed for bulk RNA-seq data, identify differentially expressed genes between two groups of samples by comparing the average gene expression between them. Although systematic evaluation of accuracy has shown these methods do not perform worse than methods that have been developed for single-cell RNA sequencing data, their accuracy depends highly on suitable filtering of the data prior to the analysis (Soneson and Robinson, 2018).

While distributions of gene expressions over bulk samples are often unimodal, single-cell gene expressions are often multimodal, even when all cells represent a homogeneous population (Korthauer et al., 2016). This is due to both biological and technical variability between cells. Methods developed specifically for single-cell RNA-seq data can account for high levels of noise and dropouts. Such methods additionally test for differences in how a gene's expression is distributed as well as differences in proportions of zeros between groups of cells, rather than just differences in average expression (Vallejos et al., 2016).

scDD (Korthauer et al., 2016) uses Bayesian modelling to test whether a gene's expression is differentially distributed under different biological conditions. The log-transformed non-zero expressions of a gene under two conditions are used to test the null hypothesis that the data arise from two equal distributions modelled as a Dirichlet process mixture of normal distributions. The zeros that are excluded from the above test are used to test for differential proportion of zeros between the

two conditions using a $\chi^2$ test. scDE (Kharchenko et al., 2014) uses raw counts and models the gene expression as a mixture of Negative Binomial distributions, for detected reads of a gene similar to methods developed for bulk RNA-seq data (Love et al. (2014), Robinson et al. (2009)), and a Poisson distribution to account for the dropout events.

Extensive comparison of various methods for differential gene expression analysis from Soneson and Robinson (2018) showed that MAST (Finak et al., 2015) outperformed other methods developed for single-cell RNA-sequencing data. Details of this method will be discussed in Chapter 2, where MAST was selected for differential gene expression analysis.

## 1.3 Computational challenges of scRNA-seq data analysis

### 1.3.1 Technical variability and dropouts

Single-cell RNA-seq analysis poses several computational and statistical challenges (Stegle et al., 2015). Unlike bulk RNA-sequencing data, the low amounts of mRNA available and the amplification bias lead to unbalanced relative gene expressions between the cells (Kharchenko et al., 2014). Technical variability between cells causes expression levels of genes being overdispersed. Additionally, for a high number of genes no corresponding reads are observed in many cells (dropout events). Even if two cells are of the same cell type, many genes expressed in one of the cells are not detected in the other cell (Kharchenko et al., 2014). The resulting high levels of dropouts and outliers pose challenges for normalisation and use of traditional similarity metrics used in clustering of bulk RNA-seq data (Kharchenko et al., 2014).

The variance measured reflects both the biological and the technical variation, which can be split into technical noise due to sampling and technical noise due to library depth differences between cells. While the latter can be corrected by normalizing the data as explained above, the sampling noise depends on the gene expression, with noise increasing with decreased gene expression (Grün et al., 2014). It has been shown that as the expression of a gene ($x$-axis) increases, the percentage of zero values across the cells ($y$-axis) decreases (Bacher and Kendziorski, 2016). While this is observed in both bulk and single-cell data, moderately expressed genes that are captured in bulk experiments have high dropout rates in single cell RNA-seq.

### 1.3.2 Batch effect

An additional challenge is technical variability that exists between data that have been processed separately, known as a batch effect. Data that have been generated in different labs or at different times show systematic technical differences that confound biological variation of gene expression. Such differences in single cell data can lead to both differences in the dynamic range of gene expression, as well as to differences in dropout rates.

Batch effect correction methods developed for bulk RNA-sequencing data (Wu et al. (2015), Johnson et al. (2007)) assume samples within each batch are biological replicates, thus they have similar cell composition and normally distributed gene expression. Systematic differences in average gene expression between samples of different batches are identified as technical variation due to batch and regressed out.

Single cell RNA-sequencing data, on the other hand, consists of a heterogeneous mixture of cells. A single cell theoretically does not have a replicate. This absence of biological replicates poses computational challenges in integrating data that have been generated separately using existing methods.

However, data integration is important for several reasons. Increase of cells representing a population leads to more accurate results in unsupervised clustering and provides computational power to resolve transcriptionally similar and rare cell types (Kiselev et al., 2019). The vast amount of publicly available data (Regev et al., 2017) enables such analysis, assuming that it is able to overcome batch effects. Finally, in order to understand the role and altered states of cell populations in disease, cross-condition comparison is required.

## 1.4 Aims and outline of thesis

During my PhD, I worked towards the development of computational methods and pipelines for the analysis of single-cell RNA-sequencing data to enable the study of immune cell populations and their states and role within the tumour microenvironment. My main interest was to identify novel subpopulations, cell state changes across conditions that are associated with survival and immunotherapy response, as well as to discover biomarkers that could be potentially used for targeted therapies. In this thesis I present the results from this work.

The first challenge I encountered for the comparison of equivalent populations across datasets and conditions was the presence of batch effects. In **Chapter 2** I discuss the effect of technical variance due to batch in combining multiple datasets and present a new method, called scID, that allows identification of transcriptionally equivalent cell populations between datasets. Through extensive validation with several datasets and comparison to other existing methods, I show that scID is outperforming other methods in cases of datasets with strong imbalances in numbers of cells and sequencing depth.

In **Chapter 3**, I use scID and other published methods to analyse two datasets of γδ-T cells from peripheral blood mononuclear cells (PBMCs) of three healthy donors and two datasets of immune cells from two breast tumour patients. My aim is to identify what subpopulations exist, what are the gene signatures that define them and what are their putative functions. I also seek to compare the identified subpopulations between Peripheral Blood Mononuclear Cell (PBMC) and breast tumour and test whether any of the subtypes correlates with clinical phenotypes, such as survival rate, metastasis or immunotherapy response.

An alternative method to overcome batch effects allowing the comparison of cell

populations across donors and conditions is to pool cells from multiple donors in a single scRNA-seq experiment, while simultaneously decreasing the per-sample library cost. Experimental methods that enable tracking of donor identity of each cell require heavy manual processing and are costly. Computational methods, on the other hand, can enable identification of donor identities of cells based on genetic mutations. Technical factors, such as sparsity and gene fragment capturing, as well as biological factors, such as cell-type-specific gene expression, can be challenging. In **Chapter 4** of this thesis I explore deep learning to demultiplex donor identities in pooled single-cell RNA-seq datasets using genetic variation.

Finally, in **Chapter 5** I discuss how these methods can improve the analysis of single-cell data and how our understanding of γδ-T cells has improved based on the results from these four datasets. Moreover, I discuss future improvements of these methods and new directions for research according to currently open challenges.

# Chapter 2

# Mapping transcriptionally equivalent cells across datasets with scID

## 2.1 Introduction

Single cell RNA-sequencing enables the study of cell type specific changes in development and disease (Regev et al., 2017). As deeply characterised, extensively validated and annotated tissue, organ and organism level atlases are increasingly being generated (Han et al., 2020), it is worthwhile to reuse such high-quality information to identify known populations of cells in new datasets obtained from equivalent tissues across different conditions in order to study condition specific changes in cell populations. Identification of condition-specific transcriptional changes of cell populations can help understand their role in disease and treatment, for example by identifying genes that can serve as biomarkers.

There are various computational challenges in comparing cells across datasets, with the most important being batch effect. Technical variation between samples that have been processed separately interferes with biological variability (Hoffman et al. (2018), Haghverdi et al. (2018), Kiselev et al. (2019)). Batch effect can lead to different dynamic range of gene expression and different dropout rates between datasets. Thus, genes that can separate cell populations in one dataset might be noisy or completely absent from another. Such technical variation is difficult to separate from biological variation, especially in cases when pairs of data include equivalent cell types but examined under different biological conditions, e.g. control and stimulated or healthy and disease. Additional challenges are introduced when there is only partial overlap between the cell types present in each dataset. Finally, the presence of transcriptionally similar cell subtypes in the dataset poses an additional challenge as biological variability between them is even lower and closer to technical noise levels.

### 2.1.1 Batch effect correction

Batch effect correction methods have been developed to disentangle the biological from the technical variance across datasets and enable combined analysis such as clustering and identification of differentially expressed genes (Hoffman et al. (2018); Haghverdi et al. (2018)). One approach for batch effect correction is to model the gene expression variation that is due to technical bias between two datasets and alter the measured gene expression values to remove it. On the other hand, alignment approaches aim to combine two datasets using common gene expression patterns in order the equivalent cell types to overlap in a transformed reduced-dimensional space but without quantifying the effect of batch in each specific gene that is measured.

### Canonical Correlation Analysis

One approach for alignment of two or more single-cell RNA-seq datasets uses canonical correlation analysis (CCA) to identify a shared correlation structure that can be used to model the common sources of variation between the datasets (Hoffman et al. (2018), Butler et al. (2018), Stuart et al. (2019)). Similar to Principal Component Analysis (PCA), CCA returns vectors that represent 'metagenes' defined as a weighted linear combination of the highly correlated genes. Next, the cells of the different datasets are aligned in this conserved low-dimensional space of CCA vectors, using dynamic time warping to account for imbalanced cluster sizes between the datasets. The aligned data can then be clustered to identify subpopulations.

Highly variable genes, i.e. genes with higher variance than expected based on their average expression, are selected for CCA. This is expected to capture cluster-specific genes that are highly expressed in one cluster and lowly expressed in the rest of the cells. However, for genes that can identify rare populations in the data both the mean expression and the variance are expected to be very close to zero, since these genes will be zero in almost all cells of the dataset, thus they might not be selected. This might lead to rare populations being misaligned. Another limitation of CCA is that rare populations that are unique to each dataset and cannot be described by the shared correlation structure, are withdrawn from any further analysis such as clustering and identification of differentially expressed genes. However, it is still very informative to know and include populations that are specific to one dataset/condition and study of rare populations is one of the strengths of single-cell RNA-sequencing.

### Mutual Nearest Neighbours

Haghverdi et al. (2018) on the other hand developed a batch effect correction

method by matching mutual nearest neighbours (MNN). Assuming there is at least one common cell population between the datasets, MNN applies cluster-specific batch effect correction. As a first step, MNN identifies equivalent cells across the two datasets using Euclidean distance. More specifically, for each cell in each dataset, MNN the k nearest cells from the other dataset based on Euclidean distance. Then two cells, each from one of the two datasets, are expected to be equivalent if they are both in each other's k nearest neighbours. For each pair of equivalent cells any systematic difference in the gene expression is regarded as a batch effect and is then removed from the data. Given the merged, batch-effect-corrected gene expression matrix of the two datasets, downstream analysis can cluster the cells to identify subpopulations.

Another assumption of MNN is that the batch effect is orthogonal to the biological variation. However, in cases where the two datasets consist of equivalent cell populations under different conditions, for example control and treatment or normal and tumour, the batch effect is confounded by actual biological differences that will be treated as technical variance and removed by this method.

Finally, MNN assumes that the technical variance is much smaller than the true biological variance. However, it is unclear whether this is the case when comparing different scRNA-sequencing methods, e.g. plate-based versus droplet-based, or datasets with strong imbalances in sequencing depth.

A general observation from using these two methods is that cluster identities of the cells differ between clustering the data separately and after batch effect correction, even for datasets with a number of cells per cell type that is sufficient to lead to stable and trustworthy clustering with unsupervised methods. This was also the case for some of the datasets analysed in this chapter (see **Figure 2.19** and **Figure 2.22**), especially when there were strong batch effects between the reference and the target datasets.

### 2.1.2 Mapping

A different approach for identifying transcriptionally equivalent cells across datasets without the need for batch effect correction is mapping. Assuming one dataset has known clusters (curated or trustworthy due to high number of cells), features (genes) can be extracted from these reference clusters and used to identify equivalent cells in other datasets. Such dataset with known cell type classification is referred to as reference. Any new dataset we seek to label using a mapping method is referred to as target.

The most intuitive method is a biomarker-based approach that is the computational equivalent of Fluorescence activated cell sorting (FACS). In this approach, we expect that each defined reference cluster has a uniquely and highly expressed gene that can be used as a biomarker to identify equivalent cells in a different (target) dataset. Although simpler than FACS, since it does not require the biomarkers to be surface proteins, there are several computational challenges. The most important of these is the dropout problem of scRNA seq data (Bacher and Kendziorski (2016); Stegle et al. (2015); Vallejos et al. (2017)). Due to the dropout events, the probability of a biomarker being detected in all cells of the respective population is very low, even in very deeply sequenced data. Thus, computational identification of such a gene that is specifically expressed in all cells of a reference cluster is challenging, leading to methods failing to label many target cells, especially in ultra-sparse datasets.

This situation becomes even more challenging when similar subpopulations exist in the dataset. In such cases biomarkers are not categorical, i.e. they don't have distinct absence/presence state in the cells but they might be shared or expressed at lower levels. In order to distinguish between transcriptionally similar cell populations a longer list of genes is required. Classifying cells based on the collective expression of a set of genes requires, however, a more systematic

approach that can take into account the probability of gene being dropped out and its biological significance for defining a cell type.

**scmap**

To map cells from target scRNA-seq data to a reference data, scmap (Kiselev et al., 2018) extracts features from each reference cluster and uses a combination of distance and correlation-based metrics to quantify the closeness between each reference cluster's centroid and the cell in the target data and then to assign target cells to reference clusters. To reduce the effect of scale difference in gene expression between the reference and the target data, scmap uses cosine similarity and correlation metrics that range [-1, 1]. However, distance metrics in high dimensional space may not work as expected, as the contrast between the distances to different data points does not exist (Aggarwal et al., 2001).

In scmap, features are extracted based on the relationship between the average expression and the dropout rate as measured by the number of zeros. More specifically, a linear model is fitted between the mean expression and the number of zeros of each gene and genes that have higher number of zeros than expected based on their average expression are selected as cluster-specific features (Andrews and Hemberg, 2019). Markers of rare clusters, however, are expected to have high number of zeros and low mean expression, since zeros are included in the calculation, which will be the same as the behaviour of noisy genes. It is thus expected that genes of rare clusters will not be included in the extracted features and that small clusters will thus be missed.

**CaSTLe**

CaSTLe (Lieberman et al., 2018) classifies target cells using gradient tree boosting (XGBoost), a machine learning classification method that consists of an ensemble of regression trees (Chen and Guestrin, 2016) whose results are combined to

provide the final classification. Features are selected based on average expression and mutual information with the reference cluster identities and highly correlated features, with Pearson correlations greater than 0.9, being removed. The XGBoost classifier is then trained using the selected features and the reference data and subsequently used to classify the target cells.

CaSTLe was shown to outperform majority vote and linear regression classifiers but has not been assessed systematically against published single-cell RNA-seq mapping methods. Feature selection using highly expressed genes is expected to miss markers of rare clusters but this could potentially be corrected by including genes with high mutual information with the reference cluster identities. On the other hand, using both reference and target datasets combined to calculate the average expression is expected to perform poorly when pairs of data have different sequencing depths. Finally, CaSTLe arbitrarily selects without justification, at maximum, 200 features; the union of the top 100 highly expressed genes and the top 100 based on mutual information, which might overlap without any justification. However, the number of features required for a multi-class classification is expected to depend on the number of classes and the similarity between them. Consequently, 200 features might not always suffice.

### 2.1.3 Aim of this Chapter

In this chapter I present a new method called scID for identifying transcriptionally related groups of cells across datasets by exploiting information from both the reference and the target datasets without making any assumptions regarding the nature of technical and biological variation in these datasets. Unlike the previously mentioned mapping methods, i.e. scmap and CaSTLe, that use highly variable and highly expressed genes, scID uses cluster-specific differentially expressed genes extracted from the reference to label the cells in the target dataset. Alternatively, scID can use any user-specified list of genes that can

identify a population, for example from bulk RNA-seq data or manually curated, and identify cells enriched for that signature in any target dataset (**Figure 2.1**).

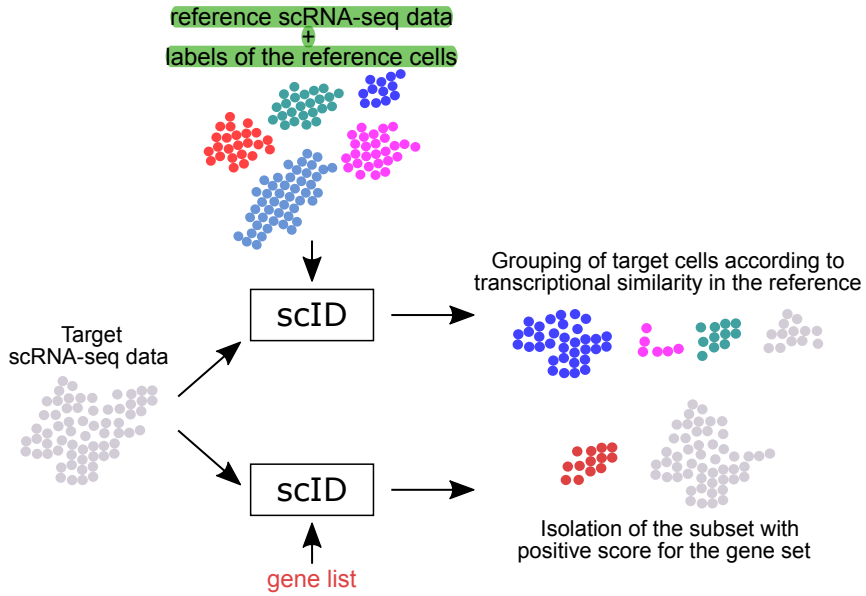



**Figure 2.1:** Graphical abstract of scID. scID can label the cells of a given scRNA-seq target dataset either based on the labels of a given labelled reference scRNA-seq dataset or identify cells enriched for a given list of genes.

I provide an extensive analysis of published datasets with batch effect and strong asymmetry in cell number and library sizes per cell for which the independent clustering of the target data via an unsupervised method is not obviously similar to that of the reference. Through this analysis I show that scID has increased classification accuracy compared to the above-mentioned alignment and mapping methods, i.e. CaSTLe, CCA, MNN and scmap. Thus, scID helps uncover hidden biological variation present between scRNA-seq datasets that often vary greatly in batch effect and quality (i.e. different numbers of cells, dropout levels and dynamic ranges of gene expression).

## 2.2 Methods

### 2.2.1 scID

Given a reference and a target gene expression matrix with rows representing genes and columns cells, and the cluster identities of the reference cells, scID identifies target cells equivalent to the reference clusters in 3 steps (**Figure 2.2**). scID splits this multiclass labelling problem into multiple binary classification problems, one for each reference cluster. Thus, for a reference cluster $C = \{r_i | L(r_i) = C\}$ consisting of all cells $r$ with class label $L()$ equal to $C$ , in the first step, scID identifies a set of $k = k_p + k_n$ features/genes (signature) that are positive ($k_p$) and negative ($k_n$) markers for this cluster, hereon referred to as a gene signature. In the second step a weight is assigned to each gene of the signature ($w_i$, for $i = 1, 2, ...k$) that represents its power to discriminate the cells of cluster $C$ from all other cells in the dataset. In the last step, scID selects target cells equivalent to reference cluster $C$ based on their score $s_j^C$. The term "equivalent" here on refers to transcriptionally equivalent cell populations with respect to the given gene signature. Given a homogeneous reference cell population with specific functions defined by the extracted gene signature, the target cells selected by scID are expected to be also biologically equivalent.

The scID score $s_j^C$ of a target cell $j$ for a reference cluster $C$ is a weighted linear sum given by:

$$s_j^C = \frac{\sum_{i=1}^{k} w_i^C \tilde{g}_i^j}{\sqrt{\sum_{i=1}^{k} w_i^{C2}}} \tag{2.1}$$

for $j = 1, ..., n$, where $n$ is the total number of cells in the target, $\tilde{g}_i^j \in [0, 1]$ is the normalised gene expression of feature $i$ in cell $j$ and $w_i^C$ is the weight that

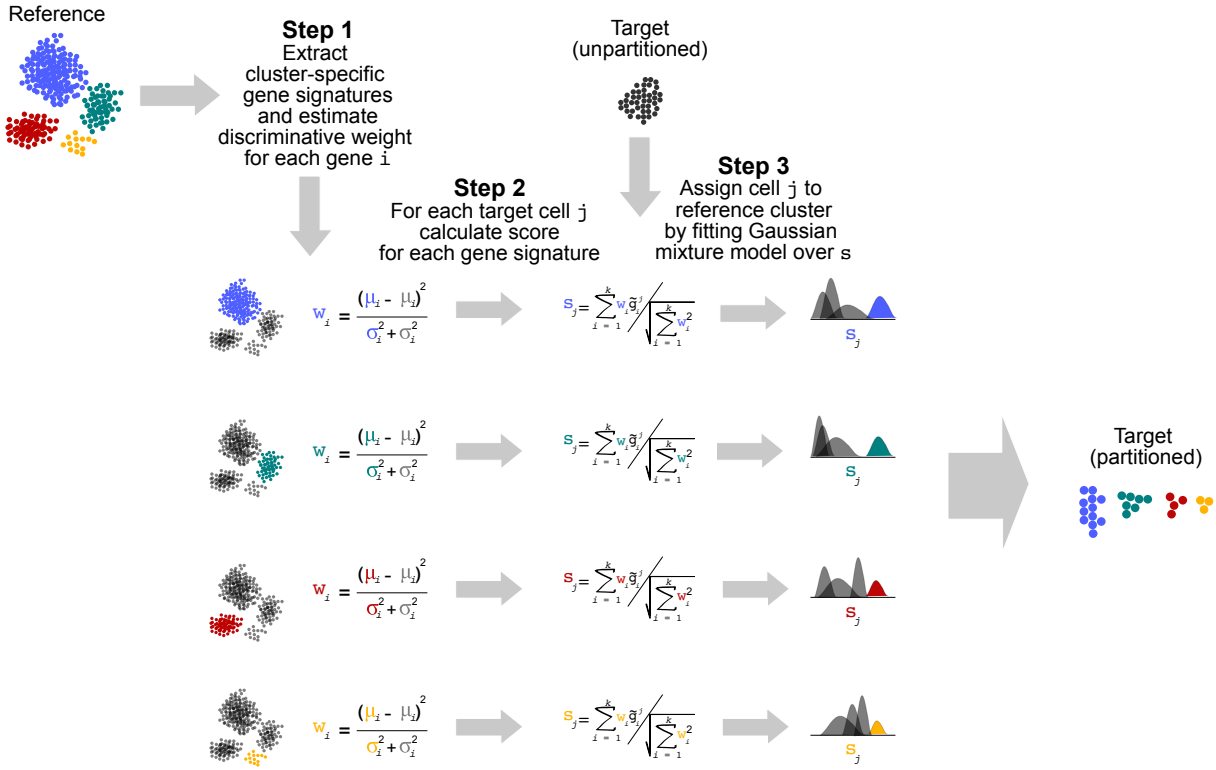



**Figure 2.2:** Overview of scID steps. The three main steps involved in mapping cells across scRNA-seq data with scID are as follows: In Step 1, gene signatures are extracted from the reference data (shown as clustered groups on a reduced dimension). In Step 2, discriminative weights are estimated from the target data for each reference cluster-specific gene signature. Finally, in Step 3, every target cell is scored for each feature and is assigned to the corresponding reference cluster.

reflects the discriminatory power of gene $i$, defined as the distance between the mean expression or the centroids of the $C$ and $C^- = \{r_i | L(r_i) \neq C\}$ populations on the projected dimension of that gene.

To reduce sensitivity of outliers and unequal contribution of genes due to differences in expression levels, the gene expression values are scaled to $[0, 1]$ by the $99^{\text{th}}$ percentile instead of the maximum, i.e. $\tilde{g}_i^j = min(\frac{g_i^j}{P_{99}}, 1)$, where $g_i^j$ is the library-depth normalised expression of gene $i$ in cell $j$ and $P_{99}$ is the $99^{\text{th}}$ percentile of its expression across all target cells. For genes that are zero in more

than 99% of the cells, the maximum can be used instead to avoid division with zero.

The weights can be computed from the reference data as follows:

$$w_i^C = \frac{\mu_i^C - \mu_i^{C^-}}{\sigma_i^{C2} + \sigma_i^{C^-2}} \tag{2.2}$$

where $\mu_i^C, \sigma_i^C$ represent the mean and standard deviation, respectively, of expression of gene $i$ in the cluster $C$; and $\mu_i^{C^-}, \sigma_i^{C^-}$ represent the mean and standard deviation, respectively, of gene $i$ in all other clusters ($C^-$). Each term of the weight is in turn calculated as follows:

$$\begin{cases} \mu_i^C = \frac{1}{l^C} \sum_{j \in C} \tilde{g_i^j} \\ \sigma_i^{C2} = \frac{1}{l^C} \sum_{j \in C} (\tilde{g_i^j} - \mu_i^C)^2 \\ \mu_i^{C^-} = \frac{1}{N - l^C} \sum_{j \in C^-} \tilde{g_i^j} \\ \sigma_i^{C^-2} = \frac{1}{N - l^C} \sum_{j \in C^-} (\tilde{g_i^j} - \mu_i^{C^-})^2 \end{cases}$$

where $l^C$ is the number of cells in cluster $C$ and $N$ is the total number of cells in the reference data.

This definition of the weights is using the framework of Linear Discriminant Analysis (LDA) (Fisher, 1936) to define a measure of the signal-to-noise ratio as the ratio of the variance between the classes to the variance within the classes (**Figure 2.3**). Based on Fisher's LDA, this metric is quantifying the distance between two classes of observations projected on a line in the direction of $w$ as:

$$w = \frac{\mu_1 - \mu_0}{\Sigma_0 + \Sigma_1} \tag{2.3}$$

An assumption of Fisher's LDA is diagonal covariance. Ideally, including the

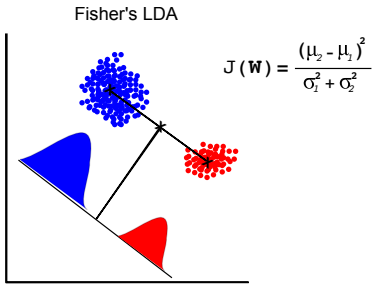



**Figure 2.3:** Fisher's Linear Discriminant Analysis for binary classification (LDA).

full covariance matrix would better capture the relationships between clusters, especially in presence of similar cell types. However, we have chosen to approximate the covariance matrix as diagonal for computational efficiency, but also due to limitations posed by the nature of scRNA-seq data. When datasets have sparse coverage, the covariance matrix is not full rank and cannot be inverted. Additionally, when the list of genes is long, computing the full covariance matrix is not just computationally inefficient but also error prone due to insufficient number of samples (cells).

Choosing the weights in this way penalises high variability and low mean expression of positive markers within the cluster of interest in order to account for the following cases:

1. When a positive marker is not expressed uniquely in the cluster of interest, $\mu_i^{C^-}$ will increase hence reducing the weight, as it does not provide sufficient evidence for classification.

2. When a positive marker is expressed only in a subpopulation of the cluster of interest, $\sigma_i^C$ will increase hence reducing the weight, so that a cell's score does not drop sharply when this gene's expression is zero. This also accounts for genes with a high dropout rate even though they might be specific and sensitive markers.

3. Finally, non-discriminative genes that are also present in other cell populations will be down-weighted, as the numerator $\mu_i^C - \mu_i^{C^-}$ will be low.

Similarly, negative markers are expected to have negative weights since $\mu_i^C < \mu_i^{C^-}$ penalizing cells that express these genes.

Although we can compute the weights from the reference data, estimating them from the target data can lead to improved accuracy, due to adjustments to the technical (e.g. library depth) and the biological (e.g cell composition) characteristics of the target data. However, to do this we need to select target cells that are likely equivalent to the reference cluster $C$. The target cells (denoted as $c$) that express the $k_p$ set of signature genes precisely and specifically and do not express the $k_n$ set of genes are selected as equivalent to the reference cluster $C$ by clustering the target cells in the differential precision - differential recall space, metrics motivated from precision and recall (**Figure 2.4**).

In general, the precision of a cell expressing a set of genes is defined as the total number of expressed genes that belong to the given set divided by the total number of expressed genes; and recall is defined as the number of expressed genes that belong to the set divided by the total number of genes in the set. So, for the set of positive markers ($k_p$)

$$\begin{cases} precision_p = \frac{n_{pme}}{n_e} \\ recall_p = \frac{n_{pme}}{n_{pm}} \end{cases}$$

where $n_{pme}$: number of positive markers expressed $n_{pm}$: total number of positive markers $n_e$: total number of expressed genes

A cell expressing all positive markers will have $recall_p = 1$ and a cell expressing only positive markers (and no other genes) will have $precision_p = 1$. However, since the majority of genes expressed in a cell will not be markers (there are housekeeping and other genes expressed), precision is expected to be lower than

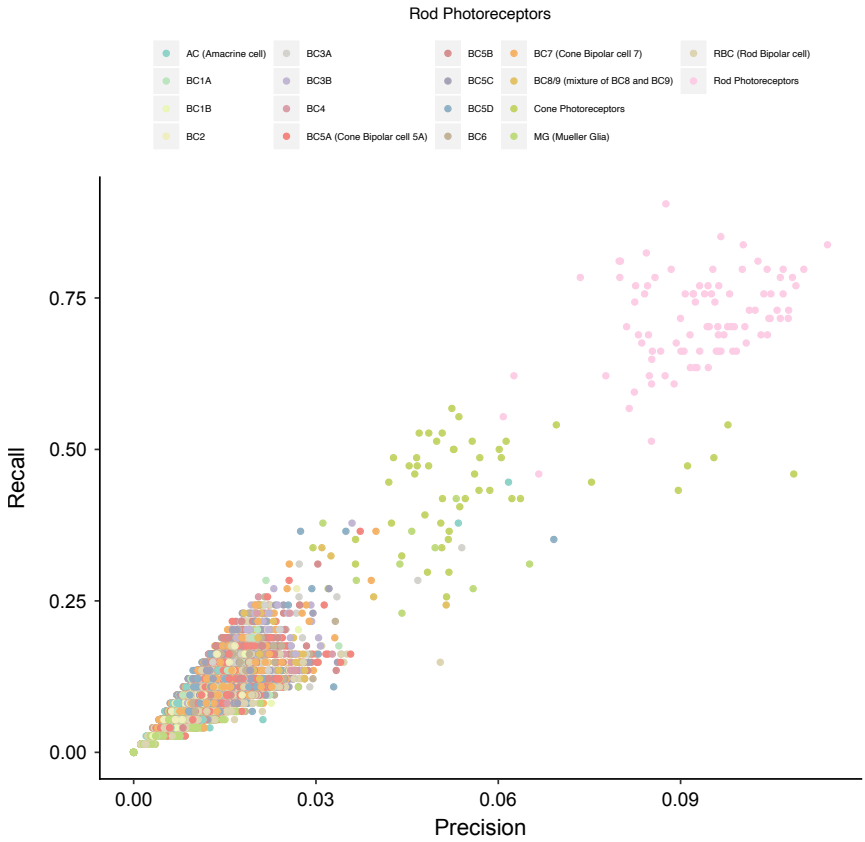



**Figure 2.4:** Projection of mouse retinal cells in the precision-recall space for the Rod Photoreceptor gene signature (Shekhar et al., 2016). Each dot represents a cell coloured based on its cell type label. Rod photoreceptors lie in the first quadrant of the precision-recall space with recall close to 1 and precision much lower than 1 but still higher than non rod photoreceptor cells.

one. Cells equivalent to the reference cluster $C$ will be in the first quadrant of the $(precision_p, recall_p)$ space, close to $(1, 1)$, and other cells not expressing positive markers will be close to $(0, 0)$ (**Figure 2.4**).

However, when transcriptionally close cell types exist in the dataset, positive markers are not sufficient to distinguish between them as some of the markers will be shared across more than one cell type (**Figure 2.5**).

For this reason, scID takes into account negative markers as well, i.e. genes that

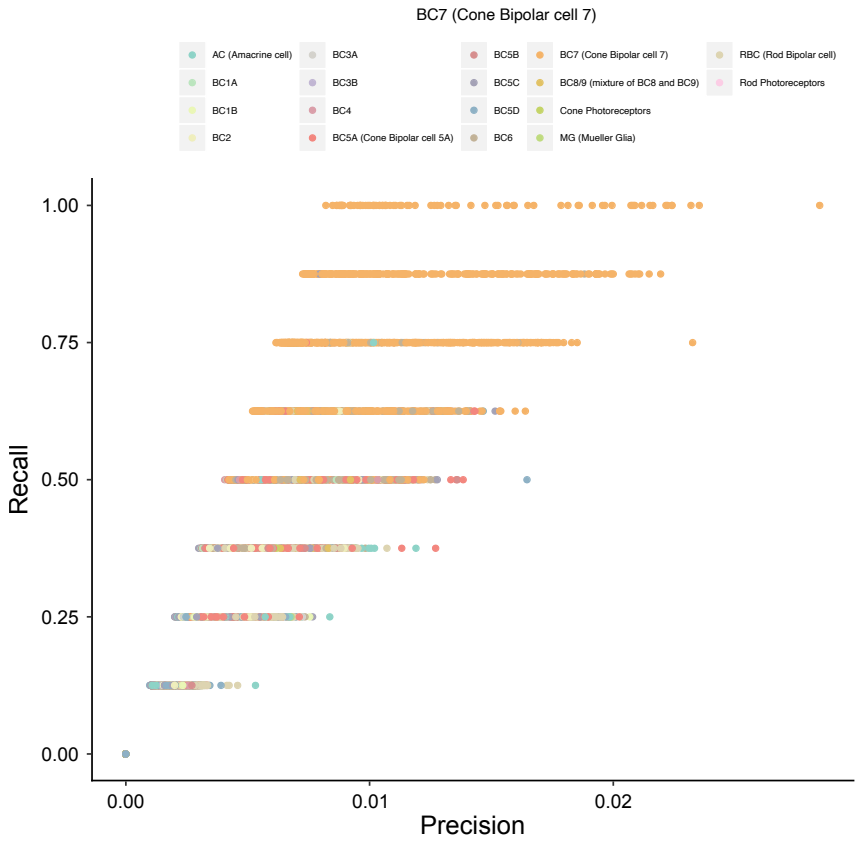



**Figure 2.5:** Projection of mouse retinal cells in the precision-recall space for the BC7 upregulated gene signature (Shekhar et al., 2016). Each dot represents a cell coloured based on its cell type label. Due to transcriptionally similar cell types being present in the dataset, BC7 cells do not cluster separately from the other cell using only positive markers..

are only present in other clusters of cells and not in $C$. Thus, for the set of negative markers $(k_n)$

$$\begin{cases} precision_n = \frac{n_{nme}}{n_e} \\ recall_n = \frac{n_{nme}}{n_{nm}} \end{cases}$$

where $n_{nme}$: number of negative markers expressed $n_{nm}$: total number of negative markers $n_e$: total number of expressed genes

Similar to precision-recall for the positive markers, a cell expressing all negative markers will have $recall_n = 1$ and a cell expressing only negative markers will have $precision = 1$. Thus, cells equivalent to the reference cluster $C$ should be in the third quadrant of the $(precision_n, recall_n)$ space, close to $(0, 0)$, while other cells expressing all or some of the negative markers will be closer to $(1, 1)$ (**Figure 2.6**).

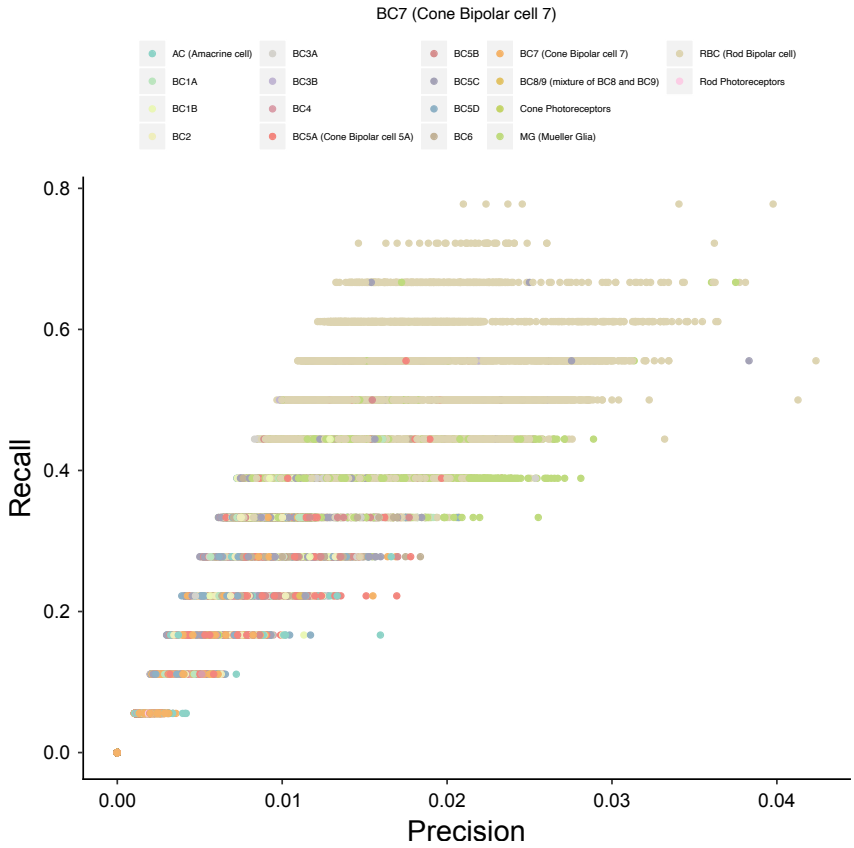



**Figure 2.6:** Projection of mouse retinal cells in the precision-recall space for the BC7 downregulated (negative) gene signature (Shekhar et al., 2016). Each dot represents a cell coloured based on its cell type label. BC7 cells are expected to be close to $(0, 0)$ since they should not express any negative markers.

To identify target cells transcriptionally equivalent to reference cluster $C$, we have combined the precision and recall for positive and negative markers into the differential precision ($DP$) and differential recall ($DR$) metrics.

$$\begin{cases} DP = precision_p - precision_n = \frac{n_{pme} - n_{nme}}{n_e} \\ DR = recall_p - recall_n = \frac{n_{pme}}{n_{pm}} - \frac{n_{nme}}{n_{nm}} \end{cases}$$

In this new space target cells equivalent to cluster $C$ will be close to $(1, 1)$, cells very different from cluster $C$ will be close to $(-1, -1)$ and cells belonging to clusters that are similar to cluster $C$ and share markers will be around $(0, 0)$ (**Figure 2.7**). This will help separate cluster $C$ from similar clusters that could be grouped together when using only the positive markers. Additionally, this will exclude doublets that consist of a cell equivalent to cluster $C$ and a cell of another cell type, as both positive and negative markers will be expressed.

To select putative positive and negative training populations from the DP-DR space, we cluster the cells using different finite Gaussian mixture models (Scrucca et al. (2016)) and select the model with the lowest Bayesian Information Criterion (BIC). The cluster with highest DP and DR is selected as putative positive cluster $c$ and the remaining cells are used as putative negative cluster $c^-$. In very rare cases where the highest DP and the highest DR refer to different clusters the cluster highest DP is selected as putative positive cluster $c$ and the cells of the cluster with highest DR are discarded from the training set. DP is expected to have higher discriminatory power than DR as it ensures that the selected set of cells has low false positives, thus providing a set of correct cells to represent cluster $c$. High DR on the other hand, ensures that all true cells belonging to $c$ are selected but does not provide any information of how contaminated the selected set of cells is with cells belonging to $c^-$.

At this point we should note that recall can be close to one for a cell that matches the gene signature in an ideal case scenario where dropout rate in the dataset is very low. Precision on the hand cannot reach 1 since a cell is expressing more genes, for example housekeeping genes, than the ones that are used for

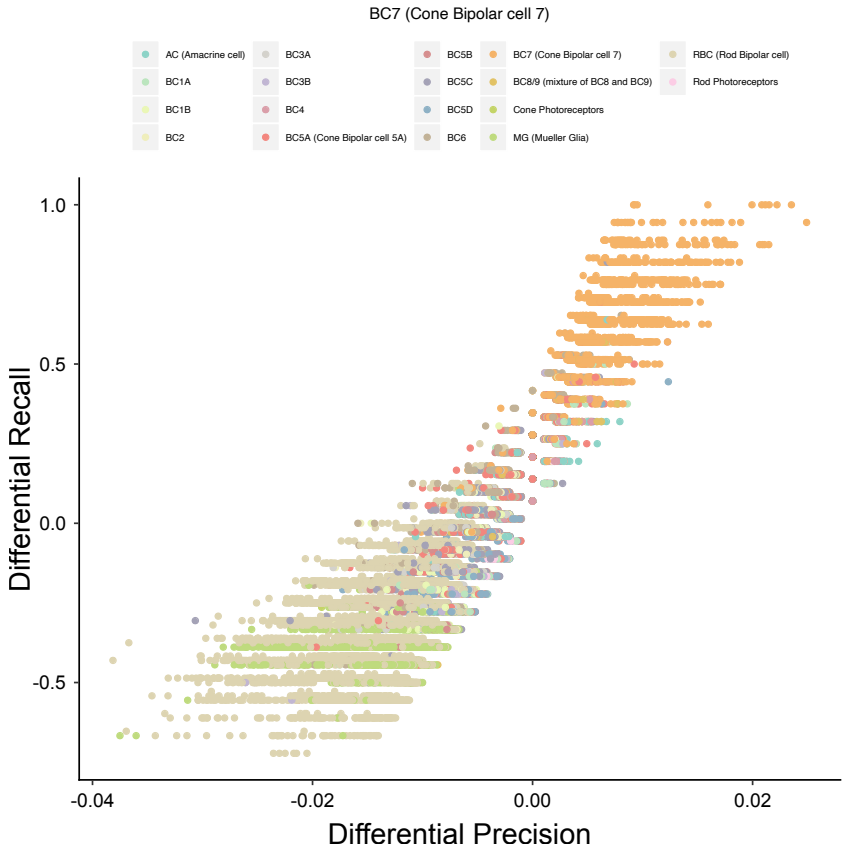



**Figure 2.7:** Projection of mouse retinal cells in the differential precision- differential recall space for the BC7 gene signature (Shekhar et al., 2016). Each dot represents a cell coloured based on its cell type label. BC7 lie in the first quadrant of the precision-recall space with recall close to 1 and precision much lower than 1 but still higher than other cell types.

classification. However, target cells that are transcriptionally equivalent to the respective reference cell type are expected to be more enriched for the signature genes and hence have higher precision than other cells. Thus clustering can be used to identify putative transcriptionally equivalent target cells.

Analogous to computing the weights from the reference data, weights can also be computed from the target dataset using the training sets of cells ($c$ and $c^{-}$).

$$w_i^c = \frac{\mu_i^c - \mu_i^{c^-}}{\sigma_i^{c2} + \sigma_i^{c^-2}} \tag{2.4}$$

where $\mu_i^c, \sigma_i^c$ represent the mean and standard deviation, respectively, of expression of gene $i$ in the cluster $c$ and $\mu_i^{c^-}, \sigma_i^{c^-}$ represent the mean and standard deviation, respectively, of gene $i$ in all other clusters ($c^-$). Each term of the weight is in turn calculated as follows:

$$\begin{cases} \mu_i^c = \frac{1}{l^c} \sum_{j \in c} \tilde{g_i^j} \\ \sigma_i^{c2} = \frac{1}{l^c} \sum_{j \in c} (\tilde{g_i^j} - \mu_i^c)^2 \\ \mu_i^{c^-} = \frac{1}{l^{c^-}} \sum_{j \in c^-} \tilde{g_i^j} \\ \sigma_i^{c^-2} = \frac{1}{l^{c^-}} \sum_{j \in c^-} (\tilde{g_i^j} - \mu_i^{c^-})^2 \end{cases}$$

where $l^c$ is the number of cells in cluster $c$.

Thus, we compute the weights from the target data, after identifying a subset of target cells ($\tilde{c}$) that likely belong to the reference cluster of interest, as follows:

$$w_i^{\tilde{c}} = \frac{\mu_i^c - \mu_i^{c^-}}{\sigma_i^{c2} + \sigma_i^{c^-2}} \tag{2.5}$$

where $\mu_i^{\tilde{c}}$, $\sigma_i^{\tilde{c}2}$ represent the mean and variance, respectively, of expression of gene $i$ in the $\tilde{c}$ set of target cells and $\mu_i^{\tilde{c}^-}$, $\sigma_i^{\tilde{c}^-2}$ represent the mean and variance of gene $i$ in the rest of the target cells $\tilde{c}^-$.

Then, to identify target cells transcriptionally equivalent to reference cluster $c$, we fit different finite mixtures of Gaussians on the scores $\mathbf{s}^C$ (Equation 2.1), select the best model as indicated by the Bayesian Information Criterion (Schwarz, 1978) and assign cells that belong to the population with highest average score to reference cluster $C$.

When the reference data consist of highly similar clusters, the features of one cluster can be correlated with the features of another cluster. This is expected to lead to scID assigning multiple reference clusters to a target cell. To resolve this, the scores of target cells $s_j^C$ are first z-score normalised and the ambiguous cells are assigned to the reference cluster with the highest normalised score. A limitation of this approach is in cases of scores of a cell for two different signatures are very close. One solution could be to explore information from the reference and define a minimum relative ratio between two close scores required for a cell to be unambiguously classified. However, this has not been implemented in the version of scID presented in this Chapter due to time limitations.

### 2.2.2 Identification of cluster-specific features

scID uses differentially expressed genes extracted from each reference cluster as cluster-specific features using the MAST method (Finak et al., 2015). MAST implements a hurdle model to identify differentially expressed genes between two groups of cells, assuming a bimodal distribution of gene expression that arises from both the stochastic nature of single cell RNA-seq data, i.e. dropout events, but also from the underlying biology of datasets that consist of transcriptionally different cell populations. Hurdle models (Cragg, 1971) are two-part models used for data with excess numbers of zero values and overdispersion and specify one process for the zero and one for the non-zero values.

In MAST, the expression level of a gene is modelled using a Bayesian logistic regression:

$$logit(P(Z_{ig} = 1)) = X_i \beta_g^D \tag{2.6}$$

where $Z_{ig}$ is a binary variable indicating whether gene $g$ is expressed in cell $i$ above zero or a selected background level.

Next a Gaussian is used to model the positive gene expression, i.e. given $Z_{ig} = 1$

as

$$P(Y_{ig} = y|Z_{ig} = 1) = N(X_i\beta_g^D, \sigma_g^2) \tag{2.7}$$

where $Y_{ig}$ is the $log_2(TPM + 1)$-normalised expression of gene $g$ in cell $i$.

Information from the two parts of the model (Eq. 2.6 and Eq. 2.7) are combined to infer cluster-specific changes in gene expression.

MAST also uses the proportion of genes detected in each cell, referred to as cellular detection rate (CDR), as a covariate in each part of the model to account for variability between cells and decrease background correlation between genes. CDR is defined as:

$$CDR_i = 1/N \sum_{g=1}^{N} Z_{ig} \tag{2.8}$$

### 2.2.3 Gene expression normalisation

When datasets were available as UMI counts, Counts Per Million mapped reads (CPM) library-depth normalisation was performed prior to the analysis with scID. CPM is a between-sample normalisation method that scales the read counts of each gene by the total number of reads captured per sample (cell) (Law et al., 2014). Thus, the CPM-normalised expression level of a gene $i$ in a cell with $N$ total reads is given by:

$$CPM_i = 10^6 \frac{X_i}{N} \tag{2.9}$$

This normalisation enables comparison of a gene across multiple cells but not comparison of expression levels of different genes within the same cell.

### 2.2.4 scID implementation

scID is available as an R library on GitHub (`https://batadalab.github.io/scID`). scID uses `MAST` as implemented in the `Seurat` package for feature extraction. Selection of training populations from the target data in the differential precision - differential recall space is implemented using the `Mclust` function of the `mclust` package and the final identification of equivalent cells based on the scID scores is implemented using the `densityMclust` function from the same package.

scID has the following user-specified options:

- `logFC`: Threshold of $log_e$ fold change between the mean expression of a gene in the cluster of interest $C$ and the mean expression of that gene in all other cells $C^{-1}$ for extracting cluster-specific genesets from the reference data.
- `estimate_weights_from_target`: Option to estimate the gene signature weights using the target data by selecting training sets using the precision-recall-like approach (default).
- `only_pos`: Use only positive marker genes from each reference cluster. When set to FALSE both positive and negative markers are used.

### 2.2.5 Biomarker-based classification of cells

To assign labels to target cells using the biomarker-based approach we implemented a computational method equivalent to FACS. To overcome missing cells due to dropouts we used two markers for each cell-type instead of one (referred to as biomarkers). Using more than two markers increases the chance of observing one of the biomarkers in a cell of a different cell-type, hence increasing the number of ambiguous cells, as explained below. The two genes with highest log

fold expression change from differential expression analysis are used per reference cluster.

Such biomarkers are expected to be highly expressed and not just present in the respective cell type. Thus we need to select a gene-specific threshold of expression over which we accept that the biomarker is highly expressed. Since expression level is relative to each dataset, depending on the sequencing depth and other technical characteristics, we infer this threshold from the expression of the gene in the target dataset. To be unbiased, we select thresholds based on different fractions of the gene's expression across all target cells (0.10, 0.25, 0.50 and 0.75). For thresholds below 0.1 all cells were classified as ambiguous and for thresholds over 0.75 all cells were classified as orphans. Using one of these thresholds of gene expression, we binarize the gene expression data, and count which biomarkers are present (value of 1) in each cell.

We require at least one of the biomarkers of a cell type to be expressed in the cell for labelling. Cells that express biomarkers of a single cell type are assigned to the respective cell type. Cells that express biomarkers of multiple cell types are labelled as ambiguous and cells that do not express any biomarker are labelled as orphans. For assessing the performance of alignment and mapping methods when "ground truth" labels are not available, we use the uniquely classified cells from this approach.

### 2.2.6 Evaluation of classification accuracy

To evaluate the classification accuracy of each method we used the following evaluation metrics.

### True Positive and False Positive Rate

In binary classification, the True Positive Rate (TPR), also referred to as sensitivity or recall, is used to measure the proportion of the true instances that have been correctly identified by the method (**Figure 2.8**). It is defined as:

$$TPR = \frac{TP}{TP + FN} \tag{2.10}$$

where $TP$ is the number of true instances that have been correctly identified and $FN$ is the number of true instances that have been incorrectly rejected.

The False Positive Rate (FPR), is used to measure the proportion of the true instances that have been correctly identified by the method. It is defined as:

$$FPR = \frac{FP}{FP + TN} \tag{2.11}$$

where $FP$ is the number of instances that have been incorrectly identified as positive and $TN$ is the number instances that have been correctly rejected.

### Adjusted Rand Index (ARI)

Rand index (RI) (Rand, 1971) is a metric of similarity between two multi-class partitions by measuring the number of agreements and disagreements. In cell clustering, Rand Index will be the ratio between the number of pairs of cells that have been grouped together in both classifications and the total number of possible cell pairs. The Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), that is widely used in comparing classification algorithms, is a corrected-for-chance modification of the Rand Index.

Given a set of cells $C = c_1, ..., c_m$ and two different partitions of the cells

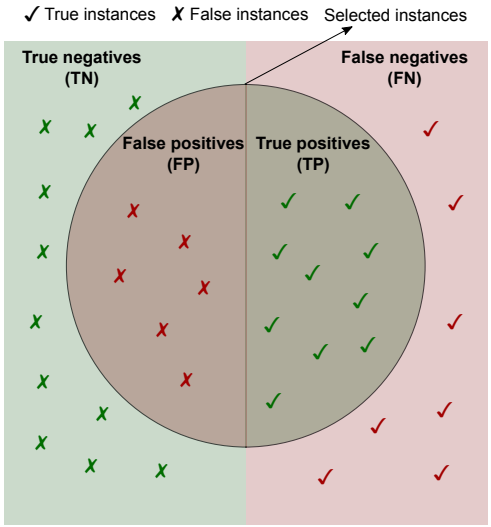



**Figure 2.8:** Illustration of the four different categories of instances of binary classification. True instances are indicated by green circles and false instances are indicated by red circles. All selected instances lie in the grey circle. True instances that lie in the grey circle are true positives (TP), true instances that lie outside the grey circle are false negatives (FN), false instances that lie in the grey circle are false positives (FP) and false instances that lie outside the grey circle are true negatives (TN).

$U = u_1, ..., u_k$ and $V = v_1, ..., v_l$, consisting of $k$ and $l$ clusters respectively, ARI is defined as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{m}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{m}{2}} \tag{2.12}$$

where $n_{ij}$ is the number of pairs of cells that belong in clusters $i = 1, ..., k$ and $j = 1, ..., l$ of the $U$ and $V$ partitions respectively, $m$ is the total number of cells in $C$ and $a_i$ and $b_j$ are the total number of cells in clusters $i = 1, ..., k$ and $j = 1, ..., l$ of the $U$ and $V$ partitions respectively. All values can be obtained from the contingency Table 2.1.

**Table 2.1:** Contingency table for calculating the adjusted rand index (ARI) between two partitions.

| | $V_1$ | $V_2$ | ... | $V_l$ | **Sums** |
|---|---|---|---|---|---|
| $U_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1l}$ | $a_1$ |
| $U_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2l}$ | $a_2$ |
| ... | ... | ... | ... | ... | ... |
| $U_k$ | $n_{k1}$ | $n_{k2}$ | ... | $n_{kl}$ | $a_k$ |
| **Sums** | $b_1$ | $b_2$ | ... | $b_l$ | |

The RI values range between 0 and 1, with 0 indicating complete disagreement of the two partitions for all pairs of cells and 1 indicating identical partitions. ARI, on the other hand, taking into account agreement by chance, ranges between -1 and 1. Random partitions will have ARI close to 0 and good partitions ARI close to 1. Negative ARI means that the two partitions have more than random disagreements; however, this is rarely the case.

### Variation of Information (VI)

Another metric that is widely used for evaluation of multi-class classification methods is the Variation of Information (VI) (Meil, 2007). VI is an information-based distance metric, similar to mutual-information. Given a set of cells $C = c_1, ..., c_m$ and two different partitions of the cells $U = u_1, ..., u_k$ and $V = v_1, ..., v_l$, consisting of $k$ and $l$ clusters respectively, VI is defined as:

$$VI = -\sum_{ij} r_{ij}[log(r_{ij}/p_i) + log(r_{ij}/q_j)] \tag{2.13}$$

where $p_i = |U_i|/m, i = 1, ..., k$, $q_j = |V_j|/m, j = 1, ..., l$ and $m$ is the total number of cells in $C$.

VI is non-negative and is equal to 0 when the two partitions are identical. Thus the closer to 0 the VI the better the agreement between the two partitions. The upper bound of VI depends on the number of clusters.

### 2.2.7 Quantification of batch effect between pairs of scRNA-seq data

To measure the extent of batch effect between the reference-target pairs of data used for the evaluation of scID and other methods throughout this chapter, I used kBET (Büttner et al., 2019) (version 0.99.5), that utilizes a $k$-nearest neighbour test. First, a $k$-nearest-neighbour matrix is created using all cells from both batches. Then, repeated $\chi^2$ tests are performed by selecting a random local neighbourhood and comparing its label distribution to the label distribution in the full dataset. If the distributions are not similar then the null hypothesis that the batches are well mixed is rejected. A binary result, i.e. 0 for "not rejected" and 1 for "rejected", is returned from all tests and the final kBET result is the average rejection rate. The higher the average rejection rate the greater the batch effect between the two scRNA-seq datasets.

### 2.2.8 Materials

#### Data source

Human Metastatic Melanoma immune cells from Tirosh et al. (2016): This Smart-seq2 data consists of a total of 4,645 malignant, immune and stromal cells from metastatic melanoma tumours from 19 patients. We have used the 3,254 immune (CD45+) cells for our analysis, which on average had 3,925 genes expressed per cell. Data was downloaded from the Broad Institute Single Cell Portal.

Mouse Retinal Bipolar Neurons from Shekhar et al. (2016): This study performed Drop-seq and Smart-seq2 experiments on Vsx2-GPF mouse retinal cells. The Drop-seq data had 27,499 cells with an average of 880 genes per cell. The Smart-seq2 data had 288 cells with an average of 4,556 genes expressed per cell. Gene expression data was downloaded from the Broad Institute Single Cell Portal.

Brain cells from E18 mouse: This Chromium 10X data consists of brain cells from the cortex, hippocampus and subverticular zone of an mouse at embryonic day 18 (E18). The scRNA-seq dataset had 9,128 cells with an average of 2,500 genes expressed per cell and the single nuclei RNA-seq data have 954 cells with an average of 2,832 genes expressed per cell. Both of these datasets were downloaded from 10X Genomics (`https://support.10xgenomics.com/single-cell-gene-expression/datasets`).

Murine tracheal epithelium cells from Montoro et al. (2018): This combination of plate-based (Smart-seq2) and droplet-based (Chromium 10X) scRNA-seq data of murine airway epithelial cells consists of 7,193 cells with an average of 1,712 genes expressed per cell. The cells were partitioned into seven clusters annotated post hoc using a biomarker approach by the authors. Data was downloaded from the Broad Institute Single Cell Portal.

Mouse brain cells from Hu et al. (2017): This Drop-seq single nuclei RNA-seq data from cortical tissues of adult mice consists of 18,194 cells with 1,649 genes expressed per cell on average that were partitioned into 40 annotated clusters. Data was downloaded from the Broad Institute Single Cell Portal.

Unstimulated and stimulated PBMCs from Kang et al. (2018): This Chromium 10X data consists of 14039 human PBMCs from eight patients, split into two groups; one control and one stimulated with interferon-beta (IFN-$\beta$). Seurat

CCA (Butler et al. (2018)) was used to align and cluster the data in order to obtain "gold standard" cell identities.

Human pancreatic islet cells from Segerstolpe et al. (2016): This Smart-seq2 data consists of pancreatic tissue and islets from six healthy individuals and four type 2 diabetes patients. Reads Per Kilobase of transcript per Million mapped reads (RPKM)-normalised gene expression data and cell labels were downloaded from ArrayExpress (E-MTAB-5061).

Human pancreatic islet cells from Grün et al. (2016): This CEL-seq data consists of pancreatic cells from deceased organ donors with and without type 2 diabetes. Gene expression data and cell labels were downloaded from NCBI GEO (GSE81076).

Human pancreatic islet cells from Muraro et al. (2016): This CEL-seq2 data consists of islets from cadaveric pancreas. Gene expression data and cell labels were downloaded from NCBI GEO (GSE85241).

Human lung adenocarcinoma cell lines from CellBench (Tian et al., 2019): Three cell lines HCC827, H1975 and H2228, were cultured separately and single cells from each line were mixed together in equal proportions. Three single cell RNA-sequencing data were generated using three different protocols, CEL-Seq2, Drop-seq and Chromium 10X.

**Table 2.2** summarizes the data used, providing information about the number of cells, library depth and number of cell types where available from the original publication.

**Table 2.2:** Summary table of published datasets used for validation of scID throughout this chapter.

| **Dataset** | **Description** | **Technology** | **Nr. of cells** | **Avg. Nr. of genes expressed per cell** | **Nr. of cell-types** |
|---|---|---|---|---|---|
| Segerstolpe et al. (2016) | Human pancreatic islet cells | Smart-Seq2 | 2394 | 6214 | 12 |
| Grün et al. (2016) | Human pancreatic islet cells | CEL-Seq | 1004 | 3467 | 12 |
| Muraro et al. (2016) | Human pancreatic islet cells | CEL-Seq2 | 2285 | 5275 | 12 |
| Hu et al. (2017) | Mouse brain cells | Drop-Seq | 18,194 | 1,649 | 40 |
| Montoro et al. (2018) | Murine tracheal epithelium cells | Smart-Seq2 and Chromium 10X | 7,193 | 1,712 | 7 |
| Shekhar et al. (2016) | Mouse retinal bipolar cells | Drop-Seq | 27,499 | 880 | 18 |
| Shekhar et al. (2016) | Mouse retinal bipolar cells | Smart-Seq2 | 288 | 4,556 | 18 |
| Tirosh et al. (2016) | Metastatic melanoma infiltrating immune cells | Smart-Seq2 | 3,254 | 3,925 | 7 |
| Zheng et al. (2017) | Brain cells | Chromium 10X | 9,128 | 2,500 | Not available |
| Zheng et al. (2017) | Brain cells | Chromium 10X | 954 | 2,832 | Not available |
| Tian et al. (2019) | Human lung adenocarcinoma cell lines | CEL-Seq2 | 274 | 7135 | 3 |
| Tian et al. (2019) | Human lung adenocarcinoma cell lines | Drop-seq | 225 | 5737 | 3 |
| Tian et al. (2019) | Human lung adenocarcinoma cell lines | Chromium 10X | 902 | 9,054 | 3 |

## 2.3 Results

### 2.3.1 Selection of training cells from target data

To evaluate the accuracy of precision/recall-like approach for identification of training cell populations from the target data, we measured the true positive and false positive rate of selected cells of different cell types in several published datasets. We observed that although the precision/recall-like approach was able to recover between 10-100% of true cells, the false positive rate was between 0 and 5% (**Figure 2.9 A**). This shows that although the precision/recall-like approach can be stringent thereby missing higher than desired levels of true cells, its role in the overall mapping is to recover true cells with a very low false positive rate. Having selected true matching and non-matching populations the weight estimation can be accurate even if not using all the cells of the dataset.

I have further demonstrated that the weights estimated from the precision/recall-like approach are accurate. I did so by calculating the Spearman rank correlation between weights calculated in Step 2 of scID and weights calculated using the cluster-based labels for the same data and for several gene signatures (**Figure 2.9 B**). For most of the gene signatures the correlation of weights is over 0.75 which indicates correct selection of training populations in Step 2 of scID. Lower correlations for some gene signatures could be explained by Step 2 selecting only a subset of the truly matching and non-matching cells and these are usually the cells with higher quality, i.e. lower dropout rate. This might lead to slightly different estimates of the variance of the gene expression in the two populations that can lead to an overestimation of the weight of the genes in the signature. However, this does not appear to affect the accuracy of mapping as will be seen later in **Figure 2.13**.

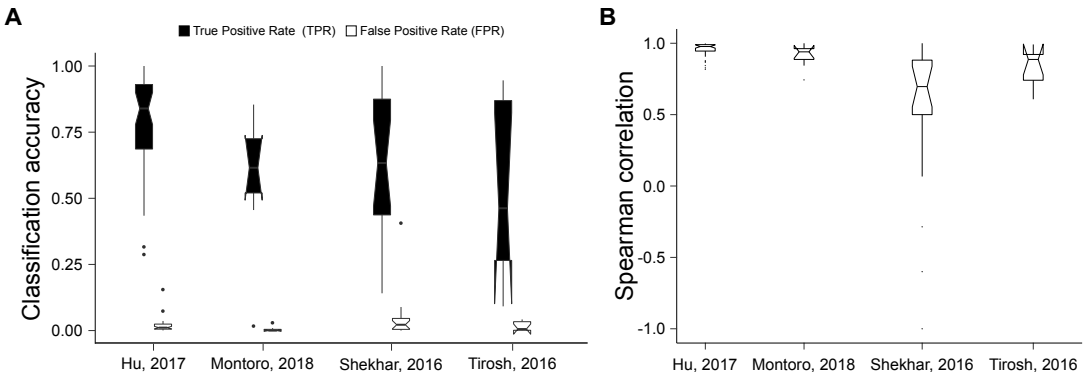



**Figure 2.9:** Quantification of accuracy of DPR classification (Step 2 of scID). **(A)** Boxplot shows interquartile range for TPR for all the cell types in each published dataset listed in the x- axis. **(B)** Boxplot shows interquartile range for FPR for all the cell types in each published dataset listed in the x- axis.

### 2.3.2 Step 3 improves putative classification of Step 2

Since Step 2 of scID might miss some cells that match the given gene signature, Step 3 aims to rescue them and correct misclassifications. Figure 2.10 shows the TPR and FPR of scID Steps 2 and 3 for each of the 18 cell-types of the mouse retinal bipolar dataset dataset from Shekhar et al. (2016). The TPR increases significantly in Step 3 over Step 2 while FPR remains low and with no significant increase, as indicated by a Kruskal-Wallis test between Step 2 and Step 3 rates.

### 2.3.3 Including negative markers improves separation between similar cell subtypes

Using positive markers, i.e. genes that are significantly upregulated in the cell type of interest compared to other cell types, is an intuitive way to label the cells, as it follows the same idea as FAC sorting. However, with single-cell RNA-sequencing being able to detect and group cells into many subpopulations with more subtle differences, sometimes similar subpopulations cannot be clearly separated using

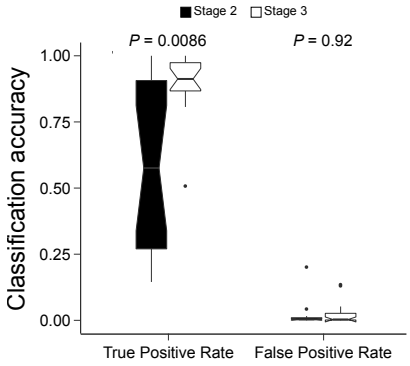



**Figure 2.10:** Quantification of TPR and FPR of Step 2 (black) and Step 3 (white) of scID. Significance was computed using a two-sided paired Kruskal-Wallis test for difference in TPR or FPR between Step 2 and Step 3.

only positive markers, especially given the sparsity of the data that has been discussed before. Figure 2.11 shows the scID scores for selected cell type gene signatures in a dataset of mouse retinal bipolar cells (Shekhar et al. (2016)). It is clear that transcriptionally distinct cell types, such as Rod Bipolar cells (RBC) and Müller Glia (MG), are easy to separate from the other cells in the dataset (**Figure 2.11 A, B**), even when using only positive markers.

However, when similar cell populations are present in the data, positive markers may not be sufficient. For example BC1B cells are transcriptionally similar to other BC subtypes, such as BC1A as indicated by the clusters' closeness in a UMAP (**Figure 2.16 A**). Using only positive markers, i.e. genes that need to be upregulated in BC1B compared to other cell types, we are unable to clearly separate this population (**Figure 2.11 C**). This is due to some genes of the signature also being expressed in these other similar cell types. However, when we include negative markers, i.e. genes that should be absent or lowly expressed in BC1B but are expressed in other cell types, we are able to increase the difference of scID scores between the true population and these near neighbours (**Figure 2.11 D**). Using negative markers the scID score can get negative values for cells that express more of the negative markers than positive markers.

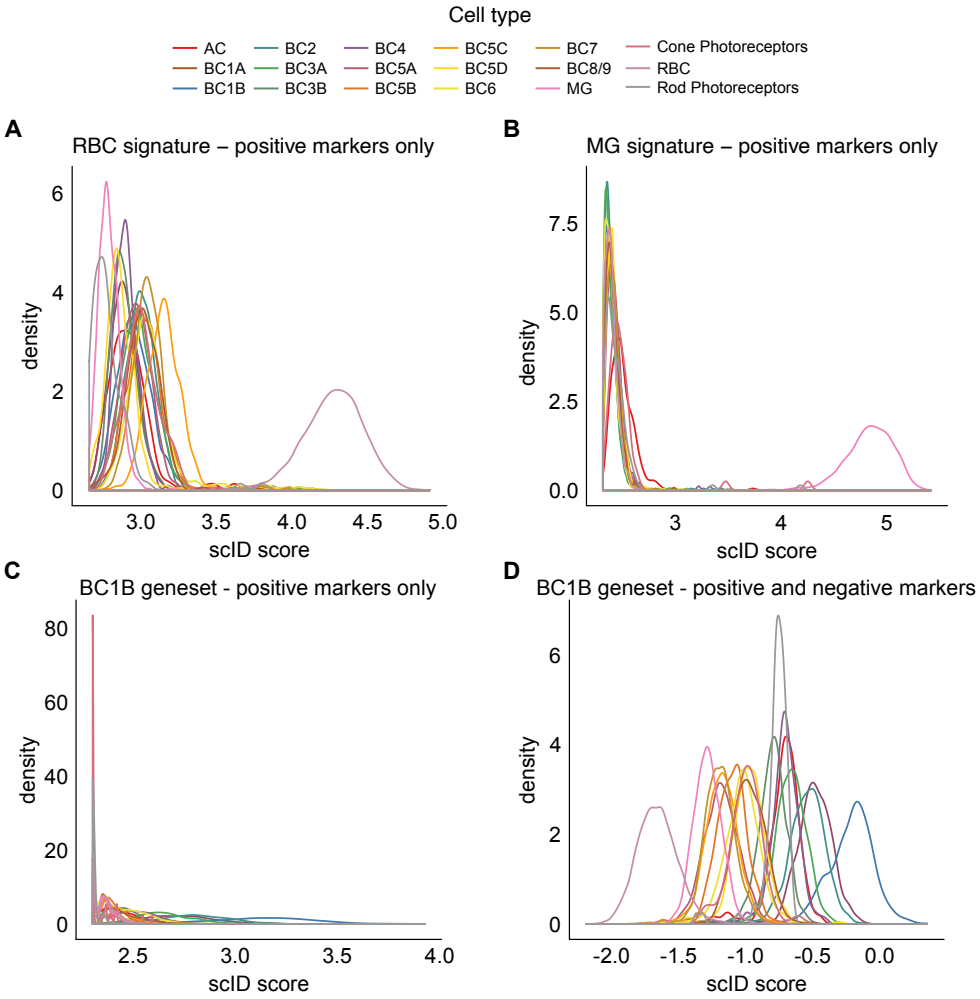



**Figure 2.11:** scID scores of mouse retinal bipolar cells (Shekhar et al. (2016)) for different gene sets using positive and negative markers. **(A)** scID scores for the RBC (Rod Bipolar Cell) gene signature including only positive markers. Since this population is abundant and transcriptionally distinct from the other cell types in the data, positive markers are sufficient to distinguish them. **(B)** scID scores for the MG (Müller Glia) gene signature including only positive markers. **(C)** scID scores for the BC1B gene signature including only positive markers. **(D)** scID scores for the BC1B gene signature including positive and negative markers.

### 2.3.4 Evaluation of scID's accuracy for datasets with ground-truth labels

As a first evaluation of scID, we tested its performance for datasets with available ground-truth labels. We obtained three datasets of human lung adenocarcinoma cell lines HCC827, H1975 and H2228 from Tian et al. (2019). These cell lines were cultured separately, then mixed in equal proportions and processed with 10X Chromium, Drop-seq and CEL-seq2 protocols, resulting in three different datasets with high batch effect (**Figure 2.12 A**). We used the 10X dataset as reference to extract cell line-specific genes (**Figure 2.12 B,C**) and mapped the Drop-seq and CEL-seq2 cells using scID. scID was able to map all CEL-Seq2 cells into the reference cell identities with high accuracy (ARI=0.98). 16 (7%) cells of the Drop-seq data were unassigned by scID, but the remaining 93% of the Drop-Seq cells were assigned to their true identities (ARI=1) as shown in **Figure 2.12 D**. Of the unassigned cells, 10 cells (62.5%) were doublets and only 6 cells were singlets.

### 2.3.5 Evaluation of scID's accuracy via self-mapping

The above datasets demonstrate scID's accuracy despite the high batch effect between the pairs of data. However, this is a simple example of having only three cell populations that are very different from each other as indicated by their distance in the t-SNE projection (**Figure 2.12 B**) and the specificity of their marker genes (**Figure 2.12 C**).

To further evaluate scID in cases of more complex datasets, we tested its performance via self-mapping for several published datasets with defined cell types. Gene signatures were extracted from published single cell RNA-seq datasets with available cell labels and scID was used to map the same cells to

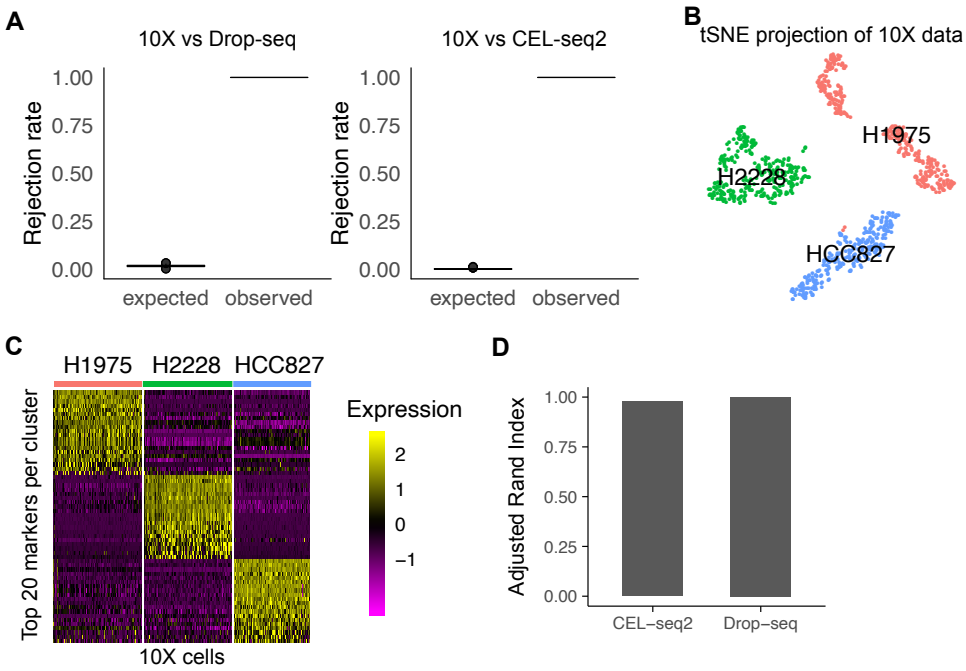



**Figure 2.12:** **(A)** Extent of batch effect between the reference (10X) and each of the target datasets (Drop-seq and CEL-seq2) as measured by kBET (Büttner et al. (2019)) is shown on the y-axis. **(B)** t-SNE projection of the cells belonging to each of the three cell lines of the 10X dataset that was used as reference. **(C)** Expression of top 20 genes (rows) specifically expressed in each cell line (columns) in the 10X data. Yellow represents enrichment and purple represents depletion of the gene's expression. **(D)** Adjusted Rand Index of scID for mapping across datasets, compared to the true labels.

the clusters using these signatures. **Figure 2.13** shows the performance of scID compared to the published labels using the Adjusted Rand Index (ARI) metric. Similarity between the scID and the true labels is very high for most of the datasets. For the datasets of Hu et al., 2017 (Hu et al., 2017), lower ARI could be explained by the high number of subtypes that exist in the dataset (**Table 2.2**), which indicates that there is higher transcriptional similarity between them that might lead to contaminated features extracted with differential expression analysis. This is a limitation of the differential expression analysis method used in Step 2 of scID in presence of many transcriptionally close cell types.

An improvement would be to refine the extracted features so that they are differentially expressed between these transcriptionally close cell types.

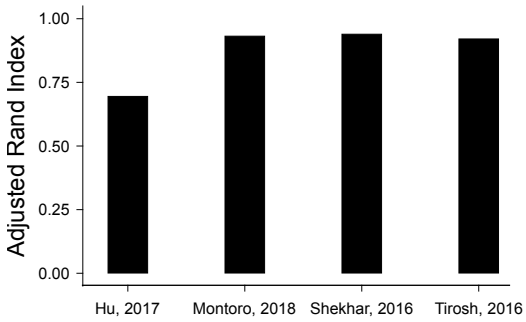



**Figure 2.13:** Assessment of accuracy of scID via self-mapping of published datasets. The indicated published data (x-axis labels) were self-mapped, i.e. used as both reference and target, by scID and the assigned labels were compared to the published cell labels.

### 2.3.6 scID is as accurate as other methods for datasets with low batch effect.

Next we compared the accuracy of scID to other alignment and mapping methods for a pair of datasets with relatively low batch effect (**Figure 2.14 A**). We used the two 10X datasets from Kang et al. (2018) for which all assessed methods are expected to be accurate due to the low batch effect. Each dataset consists of PBMC cells from the same 8 individuals before (control) and after treatment with interferon alpha (stimulated). We first co-clustered both the datasets after CCA alignment. We then applied scID and other methods, specifically MNN, scmap and CaSTLe, that were published at the time of evaluation of scID, to assign the target (stimulated) cells into cell types equivalent to the cell types of the control data, as identified by CCA. We measured the accuracy of each method by comparing the labelling from each method to the labelling in the post-CCA clustering. We can see that scID performs similarly to post-CCA clustering and CasTLe.

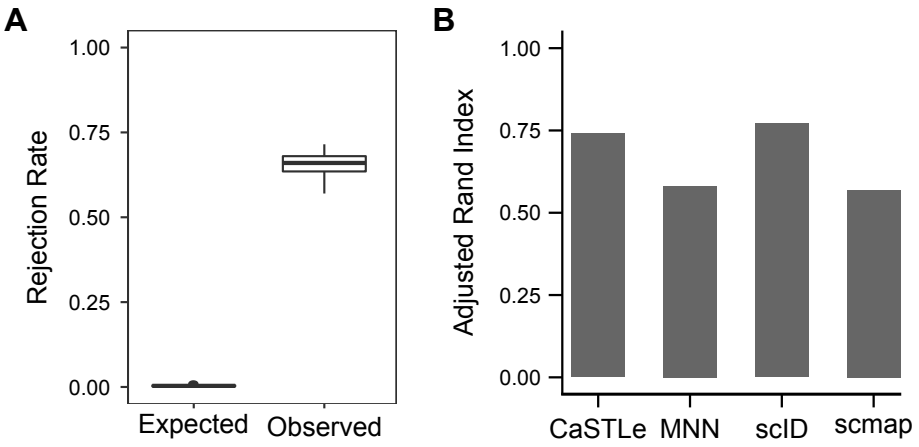



**Figure 2.14: (A)** Extent of batch effect between the reference and each of the target datasets as measured by kBET (Büttner et al. (2019)) is shown on the y-axis. **(B)** Adjusted Rand Index of scID and other methods for mapping across the two datasets, compared to the labels from CCA alignment.

### 2.3.7 Weights are adjusted to each target dataset separately

Estimation of weights from the target data enables scID to adjust to each target dataset's technical characteristics. In this example, we have three datasets of human pancreatic islets (Segerstolpe et al. (2016); Grün et al. (2016); Muraro et al. (2016)) that are expected to have similar cell-type composition. The datasets have been processed separately using three different technologies, Smart-Seq2, CEL-Seq and CEL-Seq2, that result in different cell qualities as indicated by the number of expressed genes per cell shown in Table 2.2. The scRNA-seq dataset from Segerstolpe et al. (2016) can be used as reference due to the higher number of cells and greater sequencing depth. Labels are available for all three datasets from their respective publications. I sought to classify the cells in the other two datasets into clusters equivalent to the reference cell types, using scID but also other mapping and alignment methods. We can see that both datasets have a high batch effect compared to the reference dataset (**Figure 2.15 A**). scID

and scmap are the two methods with highest performance based on the similarity of the assigned to the published labels (**Figure 2.15 B**). CaSTLe, on the other hand, has high ARI only for one of the two datasets, which indicates that the accuracy of the classification model trained on the reference data depends on the technical characteristic of the target.

The ability of scID to overcome such technical biases lies in the adjustment of gene weights for each target dataset separately. In **Figure 2.15 C** we can see the gene weights estimated from each of the target datasets, from the Grün et al. (2016) dataset on the $x$-axis and from the Muraro et al. (2016) dataset on the $y$-axis, for each reference gene set. The Spearman rank correlation of the weights for each gene set is shown in the title of each panel. We can see that the correlations diverge from 1, meaning that the scID-estimated weights for the same cell type and gene set but from different target datasets are not identical. However, scID's performance is equally high for both datasets.

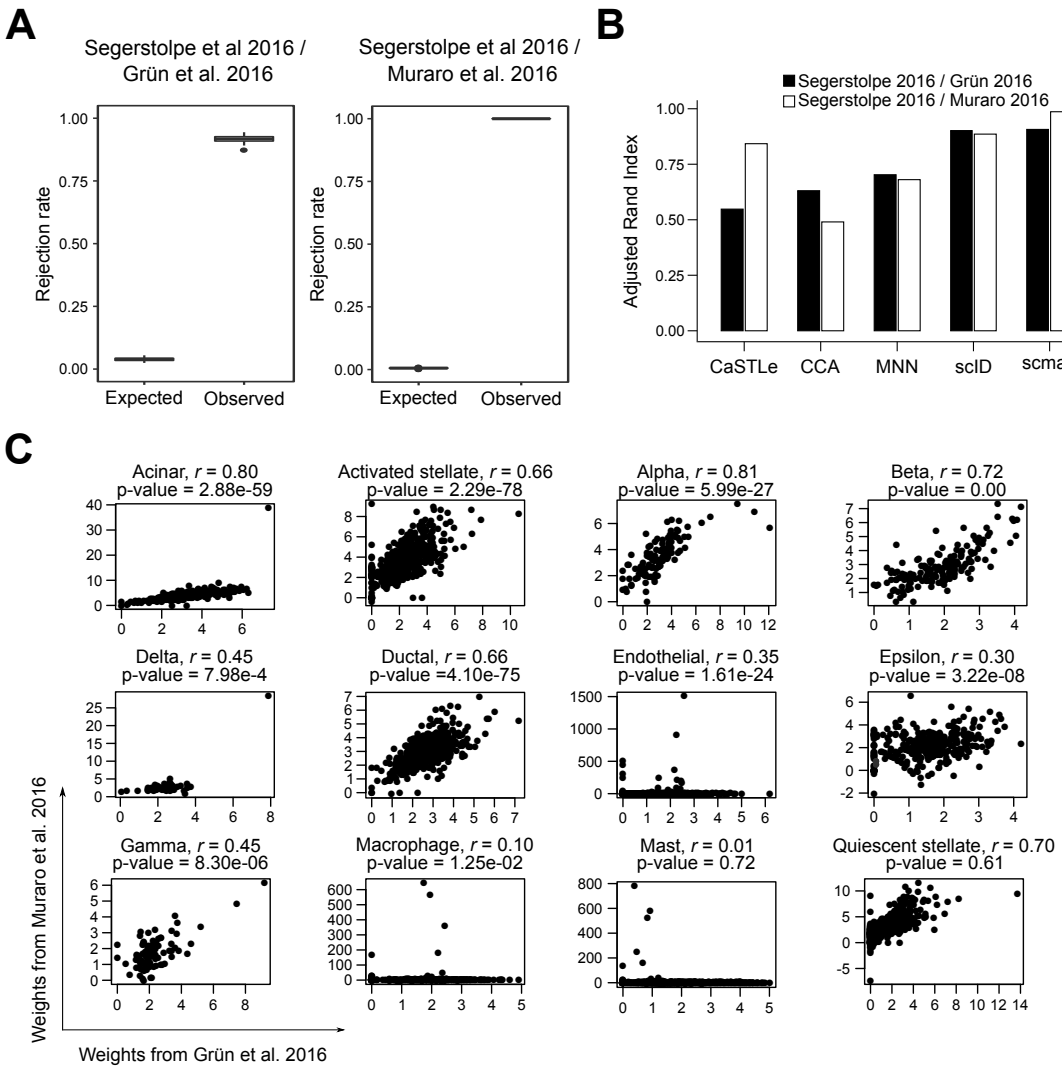



**Figure 2.15: (A)** Extent of batch effect between the reference and each of the target datasets as measured by kBET (Büttner et al 2019, Nature Methods) is shown on the y-axis. **(B)** Accuracy of final classification of scID and other methods for the two target datasets, calculated using the Adjusted Rand Index. **(C)** Scatter plot of weights estimated from pancreas scRNA-seq CEL-Seq data on the $x$-axis and weights estimated from pancreas scRNA-seq CEL-Seq2 data on the $y$-axis using pancreatic scRNA-seq SMart-Seq2 data (Segerstolpe et al., 2016) as reference. For each of the cell types in the reference data (indicated in the title of each panel), gene weights were computed using differential precision - differential recall (DPR) classification in the two target cells. Spearman rank correlation ($r$) and p-value is shown in the title of each panel. Divergence of the correlation from r =1 suggests that the weights are not identical for the two target datasets for the same cell type and gene signature.

### 2.3.8 Case Study I: Target dataset with low number of high quality cells

It is known that the accuracy of unsupervised clustering increases with the number of cells, regardless of their coverage (Kiselev et al., 2019) because even rare populations can have a sufficient number of cells to drive a separate cluster. A good example of this is the two datasets of mouse retinal bipolar cells from the study of Shekhar et al. (2016). In this study the authors have generated Drop-seq (Macosko et al., 2015) and Smart-seq2 data from the same tissue but with very different numbers of cells and library coverage; the Drop-seq data consists of 26,800 cells with around 800 genes per cell, while the Smart-seq2 data consist of 288 cells with around 4500 genes per cell. Despite the high transcriptional similarity of the cell types of the sample and the low coverage, unsupervised clustering of the Drop-seq data was able to identify 18 distinct clusters that could be annotated to known cell types (**Figure 2.16 A**). However, unsupervised clustering of the Smart-seq data resulted in only 4 clusters (**Figure 2.16 B**). As expected the batch effect between the two datasets is very strong (**Figure 2.16 C**).

One way to assess the similarity of reference and target clusters is through the expression of cluster-specific gene sets. In **Figure 2.17 A** we can see that there is a specifically enriched gene set per cluster in the Drop-Seq data. The heatmap shows the average expression of each gene set indicated in the rows, in each Drop-Seq cluster indicated in the columns, with red representing very high and blue very low expression compared to the other clusters. Using the expression of these genesets in the Smart-Seq2 clusters (**Figure 2.17 B**), we can identify Rod Bipolar Cells (RBC) and Müller Glia (MG) in clusters T0 and T3 respectively. These two cell types are both abundant and very distinct from the other cell types in the datasets, thus explaining why unsupervised clustering

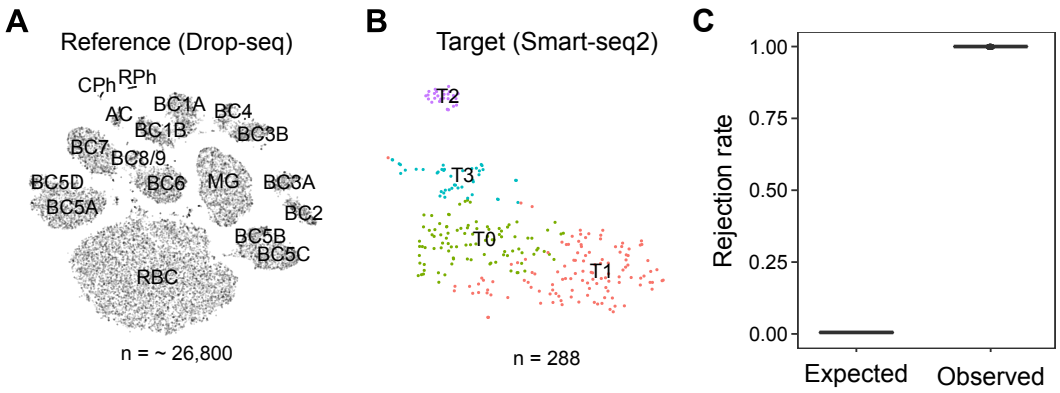



**Figure 2.16:** **(A)** t-SNE plot showing clusters in Drop-seq (reference) data of mouse retinal bipolar cells from Shekhar et al. (Shekhar et al., 2016). Cluster membership of the cells was taken from the publication. **(B)** t-SNE plot showing clusters in Smart-Seq2 (target) data of mouse retinal bipolar cells from Shekhar et al. (Shekhar et al., 2016). Data were clustered using Seurat and cluster names assigned arbitrarily. **(C)** Extent of batch effect between the reference (Drop-seq) and the target (Smart-Seq2) datasets as measured by kBET (Büttner et al., 2019) is shown on the $y$-axis.

can detect them despite their low number. The other two Smart-Seq2 clusters, however, express multiple reference gene signatures, indicating that they consist of many different cell types that due to their low number and high transcriptional similarity, unsupervised clustering did not manage to separate.

Based on the library depth of the target data, we would expect that the naïve biomarker-based approach would allow us to annotate the target cells into the reference cell types according to which markers they express. For this approach, we used the top two differentially enriched genes per reference cluster and labelled the target cells based on the expression of these two markers above a certain threshold. I used different thresholds of gene expression as explained in the Methods and counted how many different labels were assigned to each target cell. **Figure 2.18** summarises the results. It shows that only a small proportion of cells could be unambiguously classified with the biomarker-based approach.

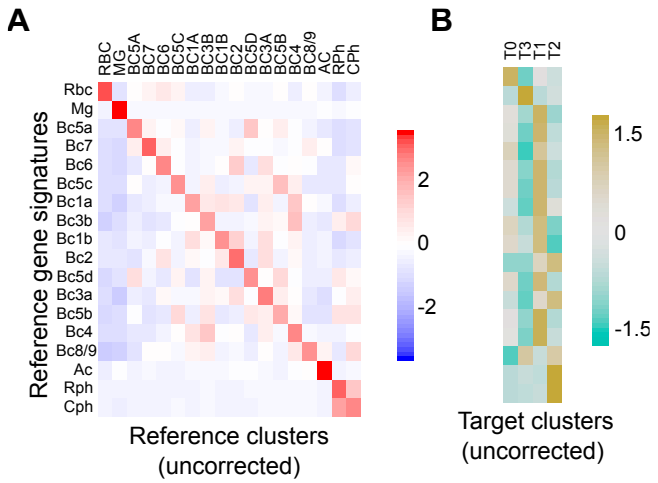



**Figure 2.17:** Heatmap showing row-scaled average expression of gene signatures (rows) in the reference Drop-seq (A) and the target Smart-seq2 (B) clusters (columns). Red (khakhi) indicates enrichment and blue (turquoise) indicates depletion of the gene signature levels relative to average expression of that gene signature across all clusters of reference (target) data.

Next I tried the combined analysis of reference and target cells after batch effect correction with CCA and MNN expecting the increased number of cells per cluster would increase the power of identification through unsupervised clustering, despite their high transcriptional similarity and low prevalence. **Figure 2.19 A** shows the heatmap of average expression of each reference gene signature in clusters after batch effect correction with CCA (left) and MNN (right). In order to be able to directly compare the result with the reference heatmap of **Figure 2.17 A**, we have retained only the reference cells in each cluster. We can see that although the Drop-Seq cells have been grouped into clusters with similar gene expression pattern as the reference data (**Figure 2.17**), there are clusters that consist of only target cells, indicating that alignment did not succeed in removing the batch effect from the two data so that equivalent cells can be grouped together. These are clusters CCA-16, CCA-18 and MNN-20 that are shown in grey as they do not contain any Drop-Seq cells. Additionally, some reference clusters have been split into multiple clusters, e.g. RBC cells have been split into two clusters (CCA-1

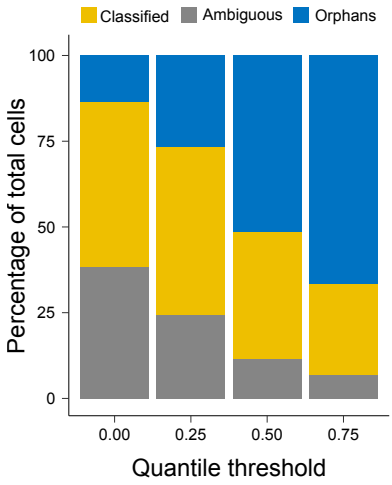



**Figure 2.18:** Identification of target (Smart-seq2) cells equivalent to reference (Drop-seq) clusters using a biomarker-based approach. Bars represent percentage of classified and unassigned cells using various thresholds for normalised gene expression (see Methods) of the marker genes as indicated on the x-axis. Gray represents the percentage of cells that express markers of multiple clusters (ambiguous); yellow represents the percentage of cells that can be unambiguously classified to a single cluster; and blue represents the percentage of cells that do not express markers of any of the clusters (orphans).

and CCA-12) after CCA and four clusters (MNN-1, MNN-2, MNN-16 and MNN-18) after MNN. Additionally, we have counted how many different signatures are enriched in each post-CCA, post-MNN and reference cluster shown in **Figure 2.19 B**. Although all reference clusters have a single gene-set enriched, many post-CCA post-MNN clusters are contaminated, expressing 2 different signatures or are not enriched for any of them, indicating random grouping of cells of different cell types.

Since alignment does not seem able to overcome the bias of batch effect due to cell number imbalances between the reference and the target data, I next tried to identify transcriptionally equivalent cell populations in the Smart-Seq2

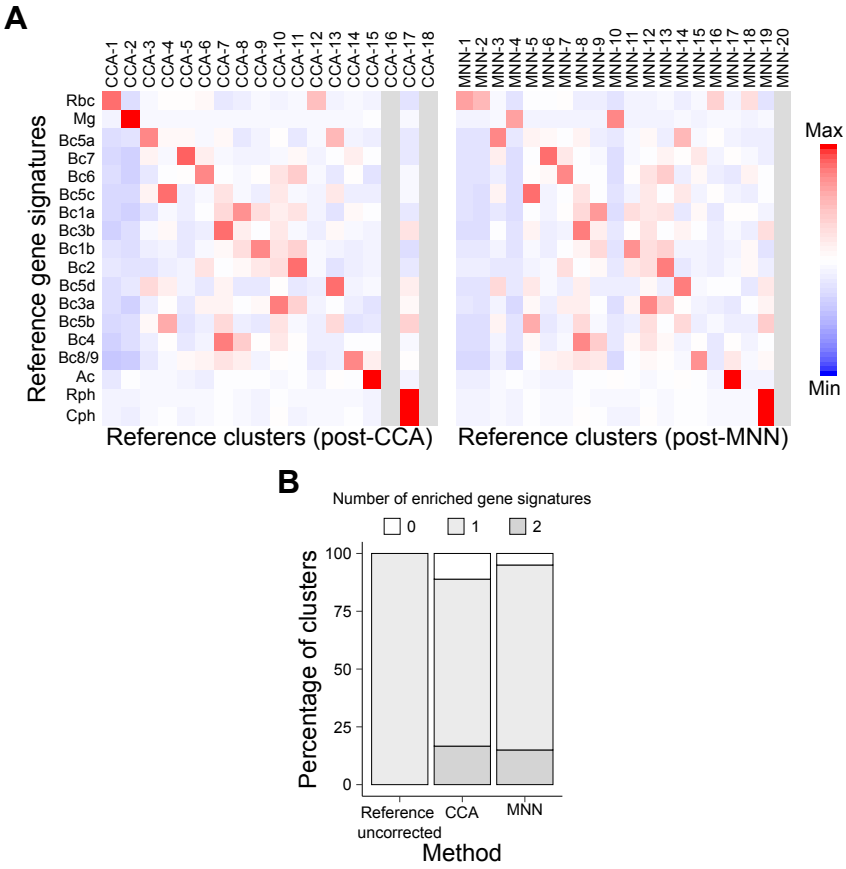



**Figure 2.19:** Batch correction by CCA and MNN alters the grouping of reference cells. **(A)** Heatmap showing row-scaled average expression of gene signatures (rows) in the reference Drop-seq clusters (columns) after batch correction of reference and target data with CCA (left) or MNN (right). Red indicates enrichment and blue indicates depletion of the gene signature levels relative to average expression of that gene signature across all clusters. **(B)** Assessment of the extent of cluster merging in post-CCA and post-MNN reference clusters compared to the uncorrected reference. Each segment represents the percentage of clusters with the indicated number of significantly expressed gene signatures.

data using mapping. scID and CaSTLe assigned 100% (n = 288) of the Smart-seq2 cells into 15 and 16 Drop-seq clusters respectively, while scmap assigned only 63.2% (n=182) of the Smart-seq2 cells into 10 clusters (**Figure 2.20 A**)

missing cells of rare clusters. Although CaSTLe seems to group the target cells in clusters equivalent to the reference data, there is very low similarity in gene expression pattern between them (**Figure 2.20 A**; right and **Figure 2.17 A**). Quantification of the number of significantly enriched gene signatures in each cluster (**Figure 2.20 B**) suggests that scmap and CasTLe have a higher proportion of merged and oversplit clusters than scID clusters. scID and scmap have the highest similarity to the biomarker-based labels for the target cells that could be successfully labelled ("Classified" category in **Figure 2.18**), while all other clusters had very low performance as indicated by the low ARI and high VI scores (**Figure 2.20 C,D**).

Thus, we can conclude that scID is outperforming other alignment and mapping approaches in datasets with imbalanced numbers of cells and cell quality by being able to more accurately identify rare populations and label cells with lower coverage.

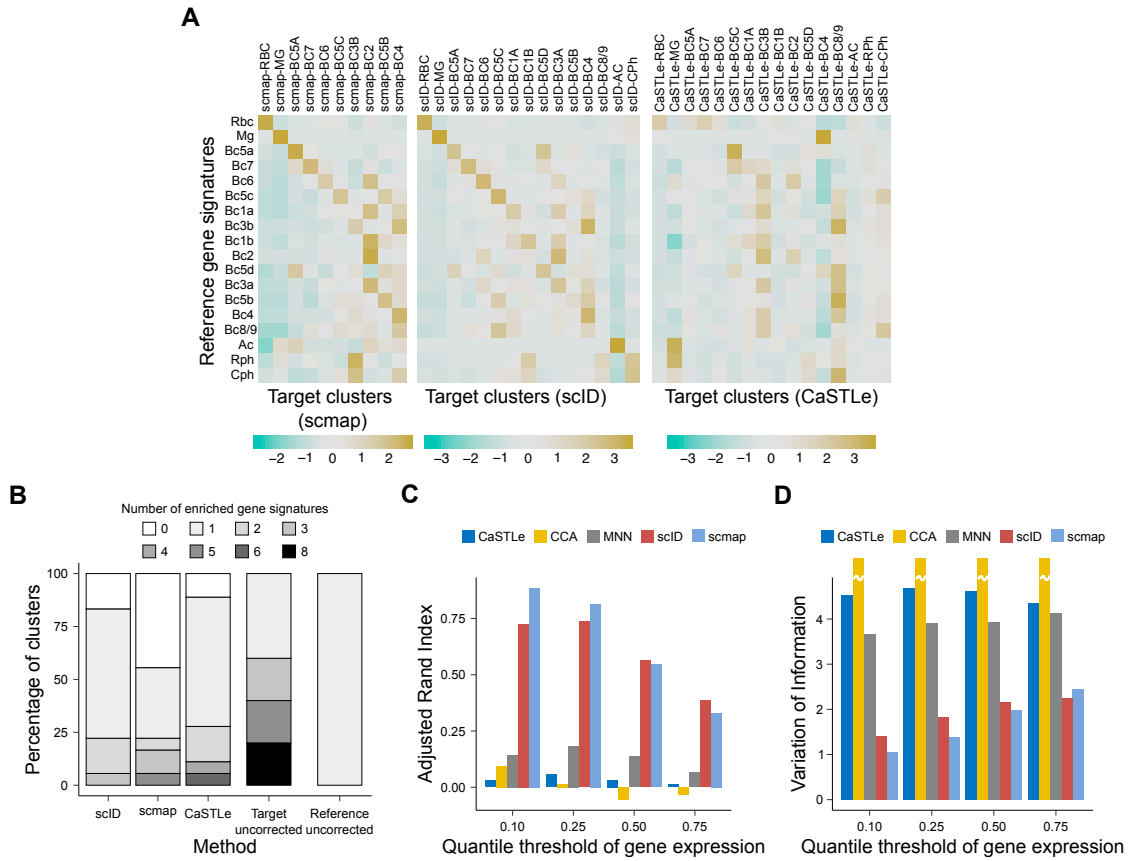



**Figure 2.20:** **(A)** Heatmap showing row-scaled average expression of gene signatures (rows) in the target (Smart-seq2) retinal bipolar data grouped by scmap (left), scID (middle) and CaSTLe (right). Khaki represents enrichment and turquoise represents depletion. **(B)** Assessment of the extent of cluster merging in scmap- and scID-mapped equivalent target cells. **(C, D)** Assessment of CaSTLe, CCA, MNN, scID and scmap. Target cells that can be unambiguously labelled using the biomarker-based approach were used as ground truth.

### 2.3.9 Case Study II: Ultra-sparse target dataset with large number of samples

In many clinical settings where only post-mortem tissues are available, single nuclei-RNA seq (snRNA-seq) (Habib et al. (2017), Lake et al. (2016)) may be possible. However, the transcript abundance in the nuclei is significantly lower than in the cytoplasm and as a consequence the complexity of scRNA-seq libraries from nuclei-scRNA-seq from the same tissue is significantly lower than in whole cell scRNA-seq (Habib et al. (2017)). These strong imbalances indicate presence of batch effect between the datasets (**Figure 2.21 B**), making it challenging to cluster even when there is a sufficiently large number of cells. Thus, we compared scID's performance to alternative approaches for a pair of publicly available Chromium 10X (Zheng et al. (2017)) based scRNA-seq data on mouse brain cells and snRNA-seq on mouse brain nuclei.

The brain whole cells scRNA-seq data had 9K cells and the brain nuclei snRNA-seq data had 1K cells. Unsupervised clustering of the data separately identified different number of clusters which did not have an obvious one-to-one correspondence (**Figure 2.21 A,C**). The higher number of cells in the whole cell scRNA-seq data is expected to lead to more accurate clustering and identification of rare clusters, thus explaining why it was used as reference for our analysis. Efforts to identify cluster membership of nuclei-seq data based on top markers in each of the clusters from the whole cell scRNA-seq data allowed unambiguous classification of only a small number of cells possibly due to shallow coverage which leads to low dynamic range in gene expression. Not surprisingly, there was a substantial number of ambiguous and orphan cells (i.e. those in which transcripts of cell markers were not captured) (**Figure 2.21 D**).

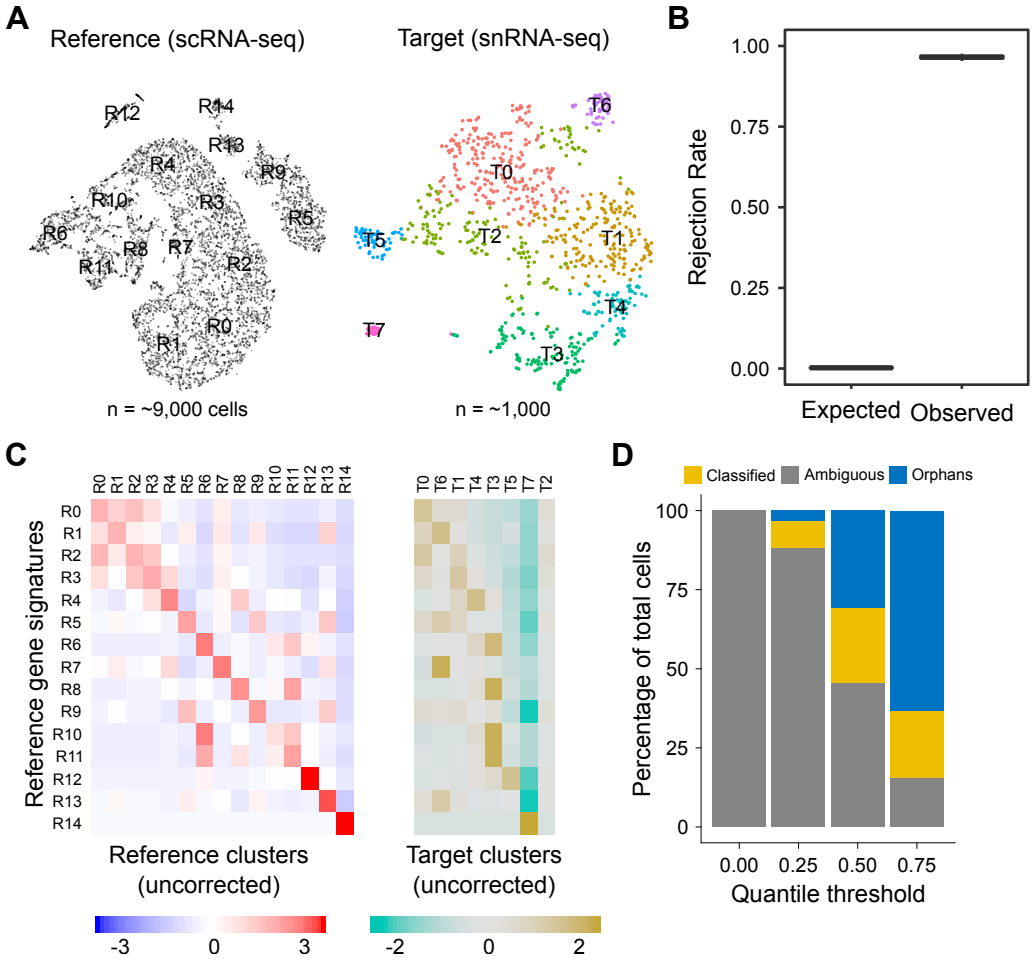



**Figure 2.21:** **(A)** t-SNE plot showing clusters in mouse brain cells (reference) and nuclei (target) data. **(B)** Extent of batch effect between the reference and the target datasets as measured by kBET (Büttner et al. (2019)) is shown on the y-axis. **(C)** Heatmap showing row-scaled average expression of gene signatures (rows) in the reference (left) and the target (right) clusters (columns). Red (khakhi) indicates enrichment and blue (turquoise) indicates depletion of the gene signature levels relative to average expression of that gene signature across all clusters of reference (target) data. **(D)** Identification of target samples equivalent to reference clusters using a biomarker-based approach obtained from the reference data.

Attempts to identify equivalent cells via joint clustering after aligning the data with either CCA or MNN appear to merge clusters of the original reference data (**Figure 2.22**).

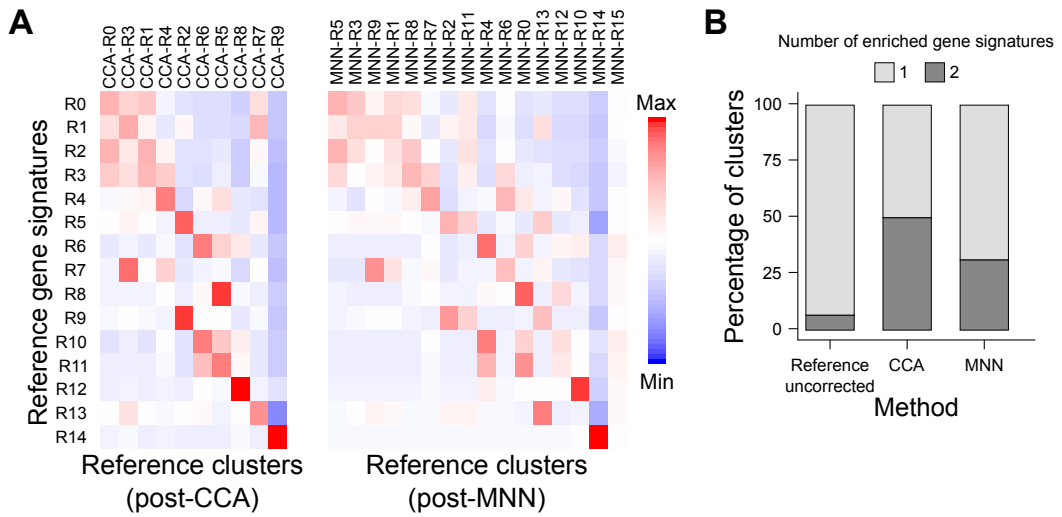



**Figure 2.22:** Batch correction by CCA and MNN alters the grouping of reference cells. **(A)** Heatmap showing row-scaled average expression of gene signatures (rows) in the reference clusters (columns) after batch correction of reference and target data with CCA (left) or MNN (right). **(B)** Assessment of the extent of cluster merging in post-CCA and post-MNN reference clusters compared to the uncorrected reference.

scID assigned 99.5% (n = 949) of the snRNA-seq cells into 14 scRNA-seq clusters while scmap assigned 60.2% (n = 574) of the snRNA-seq cells into 11 clusters and CaSTLe assigned 100% (n = 954) of the snRNA-seq cells into 15 clusters. Enrichment of reference cluster-specific gene sets in the grouped cells from all three mapping methods is similar to that in the reference data, but based on the number of enriched gene sets per group of cells for each method, scmap has higher number of merged and oversplit reference clusters (**Figure 2.23 B**). Both scmap and CaSTLe have very low similarity to the biomarker-based classification, for the cells that could be successfully annotated, while scID had the highest performance amongst all alignment and mapping methods (**Figure 2.23 C,D**).

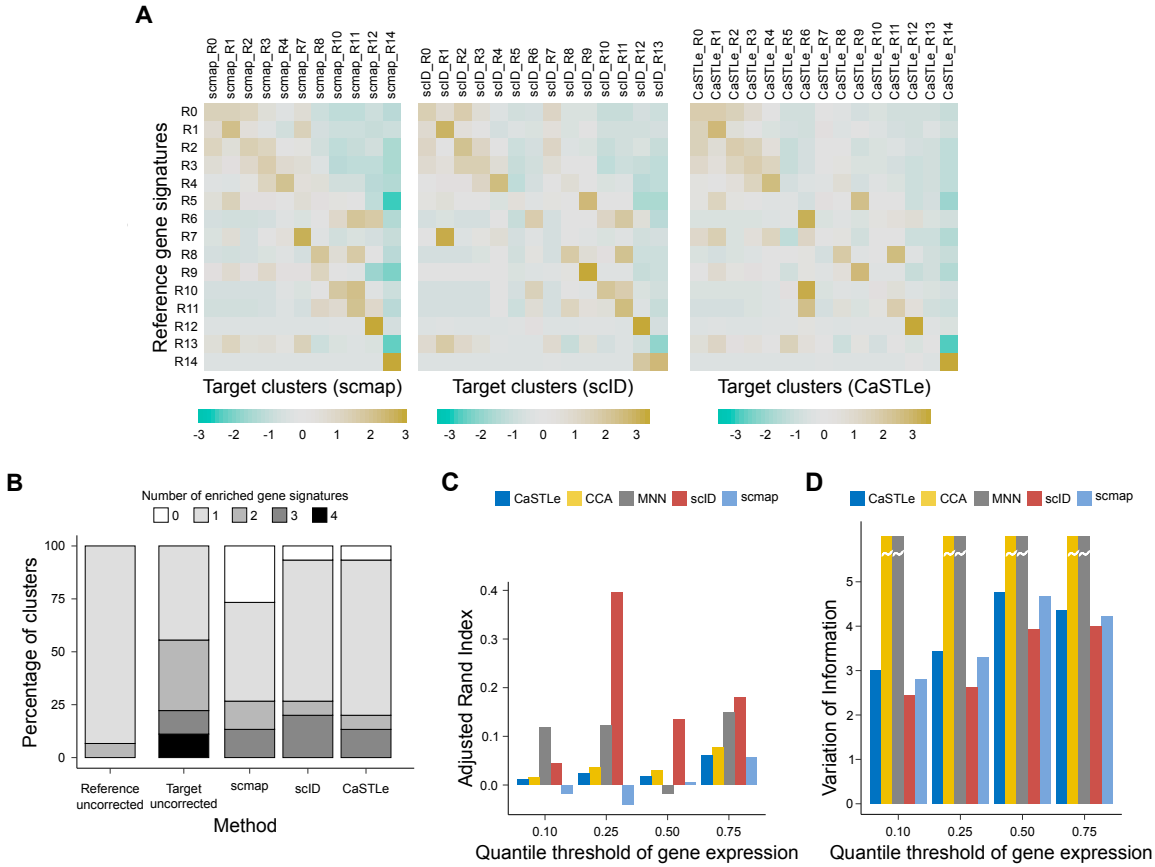



**Figure 2.23:** **(A)** Heatmap showing row-scaled average expression of gene signatures (rows) in the target data grouped by scmap (left), scID (middle) and CaSTLe (right). **(B)** Assessment of the extent of cluster merging in scmap- and scID-mapped equivalent target cells. **(C, D)** Assessment of CaSTLe, CCA, MNN, scID and scmap. Target cells that can be unambiguously labelled using the biomarker-based approach were used as ground truth.

## 2.4 Discussion

Comparison of multiple scRNA-seq datasets across different tissues, individuals and perturbations is necessary in order to reveal potential biological mechanisms underlying phenotypic diversity. However, comparison of scRNA-seq datasets is challenging because even scRNA-seq data from the same tissue but with different technology or donor can have a significant batch effect that confounds technical with biological variability. scRNA-seq datasets generated from different scRNA-seq methods and quality (i.e. similar number of cells and library depth per cell) are in general much easier to correct than data generated from different methods and quality. Particularly, in cases of strong asymmetry in quality between datasets to be compared, mapping target cells across data at the individual cell level rather than performing batch correction can be a more appropriate route for identifying known transcriptionally related cells (**Figure 2.20 C,D**; **Figure 2.23 C,D**).

scID uses the framework of Fisher's linear discriminant analysis to map cells from a target data to a reference dataset (**Figure 2.2**). Other similar linear regression methods, such as support vector machine (SVM) and logistic regression were also investigated. These two methods could accurately classify high quality datasets with very distinct cell types but performed poorly when dropout rate and transcriptional similarity between cell types in the data increased. Additionally, such methods rely heavily on the similarity of gene expression distribution between the training and the testing data and are thus not expected to perform accurately for unbalanced pairs of data with significant differences in technical characteristics, such as those presented in this chapter. On the other hand, the weights estimated by scID using the LDA framework represent the "signal-to-noise" ratio of a gene and are independent of the dynamic range (scale) of gene expression.

scID adapts to characteristics of the target data in order to learn the discriminative power of cluster-specific genes extracted from the reference in the target data. scID down-weights the genes within the signature that are not discriminatory in the target data. These genes could be either falsely selected as cluster-specific or not discriminative in the target dataset due to different data quality (e.g. they have high dropout rates) or different cell composition (i.e. presence of transcriptionally similar cell populations).

Through extensive characterisation of fourteen published scRNA-seq datasets, I assessed scID's accuracy relative to alternative alignment and mapping approaches in situations where the reference and target datasets have strong asymmetry in quality and show that scID outperforms existing methods for recovering transcriptionally related cell populations from the target data which do not cluster concordantly with the reference.

scID can either estimate gene signature weights from the reference or the target data. When the reference and the target data have a strong batch effect, estimation of weights from the target appears to be accurate, adjusting the importance of the genes to both the technical (e.g. dropout rate, dynamic range of gene expression) and the biological (e.g. tissue composition) characteristics of each input dataset. However, in cases where there are very similar rare cell types in the datasets to be compared, estimation of weights from the reference may be more suitable.

For a data set to serve as a reference, we expect well defined clusters by cell type such that cluster specific gene signatures exist and they are mutually exclusive as can be seen in **Figure 2.17 A** and **Figure 2.21 C**. As unsupervised clustering methods work better for datasets with high numbers of cells per cell type (Kiselev et al., 2019), such datasets can serve as reliable references.

However, due to a reliance on gene signatures to discriminate between clusters, a limitation of scID is that it only works for reference clusters that have distinct gene signatures. When there are very similar subpopulations in the target data with a very limited number of markers, scID may assign multiple labels to a cell and this for now has been solved by comparing the scaled scores of different matching signatures to a cell and assigning the label with the highest score.

One way to avoid multiple labels is by improving the feature extraction step. Current differential expression analysis methods compare the cluster of interest to all other cells of the dataset, forming a pooled "background". However, the signals of relatively small neighbouring clusters will be masked by the majority of the cells in the background falsely calling genes as differentially expressed. A solution to this is to identify cluster-specific genes by comparing the cluster of interest to its nearest neighbours instead of all cell population in the mixture. An alternative method to select discriminative gene sets and their weights that maximally separate a cluster of interest from its closest neighbours and thus from all other clusters in the dataset will be discussed in the following chapter.

Another limitation of scID is that if the target data contains a cell type that is only weakly related to but not identical to the cell type in the reference, scID is susceptible to falsely classifying such cells. This is due to the fact that the extracted genes are differentially expressed between the cell type of interest and only all other cell types present in the reference dataset, thus relative to the cell composition of this data. One solution to this could be to infer some minimum matching score required for a cell to be labelled from the reference data. This, however, needs to take into account the technical quality differences between the two datasets that may lead to different scales of scores. Another solution could be the improvement of extracted features by combining more than one annotated datasets from the same tissue as reference. This will enable the expansion of the cell composition of the reference data that can lead to extracted features that

discriminate the cell population of interest from all possible cell populations in that tissue. The main challenge of this is the presence of a batch effect between datasets that have been processed separately. A recently published method, SuperCT (Xie et al., 2019), uses a neural network implementation of a supervised classifier to incorporate data from the Human (Regev et al., 2017) or Mouse Cell Atlas (Han et al., 2018) for training. Such an approach could help create a big reference dataset where all known cell types are represented leading to extraction of better cell type-specific features.

### 2.4.1 Conclusion

As the scale of single cell RNA-seq data increases and the numbers of clusters obtained become so large that manual annotation is cumbersome, scID can enable automatic propagation of annotations and metadata across clusters in these datasets. Each reference dataset needs to be processed only once and then multiple target datasets can be mapped simultaneously, unlike other alignment and mapping methods that need to process the reference dataset every time a new target dataset becomes available.

scID can take advantage of the vast amount of high quality single-cell RNA-seq datasets available in public repositories such as the Human Cell Atlas (Regev et al., 2017) and the Mouse Cell Atlas (Han et al., 2018) to use as reference data for cell type annotation of new target data sets. However, unlike other methods that require reference gene expression data, scID can also be used with user-specified gene lists obtained from curation or bulk RNA-seq data. In addition, scID can be used for ordering cells based on an arbitrary user specified gene list (such as those defining a pathway) without the need for providing information on gene expression levels thus aiding in the interpretation of a subset of cells across cell types.

Future directions on enhancing scID could be to include a database of pre-extracted cell type signatures from high quality reference datasets. In this way, scID can be part of a standardised single-cell RNA-seq data analysis pipeline for cell type annotation at the cell or the cluster level. Such a database can ensure the specificity and sensitivity of cell type signatures used for the classification of a new dataset and enable the use of scID without the need for literature research to identify suitable reference datasets.

# Chapter 3

# Identification of sub-populations in circulating and breast tumour infiltrating γδ−T cells

## 3.1 Introduction

### 3.1.1 γδ-T cells in human peripheral blood

T lymphocytes can be divided into two distinct subsets, αβ-T and γδ-T cells, based on the type of chains that compose their T Cell Receptor (TCR). αβ-T cells use a CD3-associated alpha/beta (αβ) TCR for recognition of processed peptide antigens presented on Major Histocompatibility Complex (MHC). On the other hand, γδ-T cells express a gamma/delta (γδ) TCR, which can recognize both peptide and non-peptide antigens directly (Sebestyen et al., 2019).

γδ-T cells exist in both circulation and tissues (Sebestyen et al., 2019) and multiple γδ-T cell subgroups have been characterised based on their TCR composition and

cellular functions. In human four TCR δ chains (Vδ1, Vδ2, Vδ3 and Vδ5) and seven γ chains (Vγ2, Vγ3, Vγ4, Vγ5, Vγ8, Vγ9 and Vγ11) exist. Some Vδ chains preferentially pair with Vγ chains. For instance, the Vγ9-Vδ2 γδ-T subtype, which make up the majority of γδ-T cells in the peripheral blood (Dimova et al., 2015), execute surveillance of endogenous mutant cells and the eradication of exogenous invasive pathogens.

Two broad subgroups defined based on cellular function include effector γδ-T cell, which can kill cells such as tumour cells directly (Kabelitz et al., 2007), and regulatory γδ-T cell, that promote immunity through secreting cytokines (Zhao et al., 2018). Classically, γδ-T subtypes are also defined based on whether they produce IL17A or interferon gamma (IFNγ) (Ribot et al., 2009). Additionally, γδ-T cell subtypes are also divided along the innate-like and adaptive-like axis, depending on whether their cytotoxic function is induced by *NKG2D* receptor or through TCR (Davey et al., 2018). A subset of γδ-T cells, particularly the δ1 subtypes, express *NKG2D* receptor which can recognize stress-induced antigens such as MICA/B and members of the ULBP family (Groh et al., 1999). In humans, binding of stress induced ligands to NKG2D leads to activation of immune function through phosphorylation of the adaptor protein DAP10 (Billadeau et al., 2003).

Using a candidate gene approach, Ryan et al (Ryan et al., 2016) have identified *CD16* and *CD28* as markers of two distinct PBMC δ2 subtypes. However, subtypes and markers of δ1 have not been studied. Single-cell RNA-sequencing provides an opportunity to uncover γδ-T subtypes, their markers and putative functions in an unbiased way (Regev et al., 2017). Pizzolato et al (Pizzolato et al., 2019) carried out scRNA-seq of δ1 and δ2 sorted subsets of human γδ-T cells from PBMC. Although their data allowed identification of genes differentially expressed between sorted δ1 and δ2 subsets of γδ-T, the relatively low number

of cells, approx. 1200 γδ-T cells combined from multiple datasets, that were sequenced did not allow identification of subclusters within them.

### 3.1.2 The role of γδ-T cells in the tumour microenvironment

While the human γδ-T cells have been characterised largely in the peripheral blood and in the context of bacterial and viral infection, computational study of The Cancer Genome Atlas (TCGA) cancer data has shown that elevated levels of γδ-T cells in a variety of solid tumours are associated with favourable prognosis (Gentles et al. (2015), Ma et al. (2012), Wu et al. (2019)). γδ-T cells can recognize transformed cells through their engagement of their TCR and/or natural killer cell receptors (NKRs), such as *NKG2D* (Simões et al. (2018), Bauer et al. (1999)) and directly kill them through expression of TRAIL, FASL or secretion of perforin and granzyme, in a similar way as conventional cytotoxic T cells (Silva-Santos et al., 2019). Additionally, γδ-T cells can play an indirect role in antitumour immunity through the activation of αβ-T (Brandes et al. (2005), Altvater et al. (2012), (Mao et al., 2014), Himoudi et al. (2012), Muto et al. (2015)) and NK (Maniar et al., 2010) cells or the induction of class switching of B cells (Wen et al. (1996), Huang et al. (2015), Rezende et al. (2018)). Using ex vivo grid culture expansion of γδ-T cells isolated from breast tumours, Wu et al (Wu et al., 2019) identified an *IFNγ* positive innate-like δ1 subtype in breast tumour that was associated with favourable overall survival in triple negative breast cancer patients. However, markers unique to this subtype and the gene expression programs that potentially underlie their clinical association have not been defined.

γδ-T cells are also known to have pro-tumour functions. For example, increase of expression of *IL17A* has been observed in γδ-T cells in the tumour microenvironment and has been associated with tumour growth (Wakita et al. (2010), Carmi et al. (2011), Benevides et al. (2015), Kimura et al. (2016), Kulig et al. (2016), Ma et al. (2014), Rei et al. (2014), Patin et al. (2018)). *IL17A*

has several pro-tumour functions, such as angiogenesis and stimulation of tumour cell proliferation, which is also induced by expression of *IL22* (Silva-Santos et al., 2019). Additionally, γδ-T cells can inhibit DC maturation and suppress T cell responses.

### 3.1.3 Aims of this Chapter

The above diverse functions of γδ-T cells make them an interesting subpopulation for targeted cancer immunotherapy. However, we need to understand whether specific subtypes are associated with either pro- or anti-tumour functions and what are the markers that define them in order for them to be used as immunotherapy targets. Despite the above efforts to identify subtypes of γδ-T cells and understand their functions in normal and diseased tissues, their heterogeneity is not fully resolved. This is partly due to their low abundance in mammals, which makes them difficult to detect and study, but also due to the high evolutionary divergence of the TCR genes between human and mice that are often used as model organisms.

Single cell RNA-sequencing methods can enable the identification of subpopulations of γδ-T cells in an unbiased way by revealing the heterogeneity within this cell type that was masked in previous bulk RNA-seq experiments. Better characterisation of the heterogeneity of γδ-T cells can help uncover markers for their specific isolation and experimental functional characterisation.

In this chapter, I will be using four newly generated single cell RNA-seq datasets of γδ-T cells from peripheral blood of three healthy donors and breast tissue of two breast cancer patients. I will be using both alignment and mapping methods to integrated them and identify equivalent populations across the two conditions as well as other further downstream analysis of the resulting subpopulations to identify markers that characterise them and computationally

annotate their putative functions. In addition, the methods presented here can serve as an example of a standard pipeline of analysis of single-cell RNA-sequencing data, starting from alignment of sequence reads and including quality control and filtering of cells, batch effect correction through alignment and mapping, clustering, identification of differentially expressed genes and functional annotation.

## 3.2 Materials and Methods

### 3.2.1 Experimental Methods

#### Sample acquisition, processing and sorting

Breast cancer patient samples (here on referred to with prefix "BC") and blood from healthy donors (here on referred to with prefix "HD") were obtained with informed consent (NHS Lothian, Tissue Request No. 2017/SR865). Acquired samples were processed by Dr Victor González-Huici. Within approximately 30 minutes of obtaining each donor's blood, PBMCs were isolated via the Ficol density gradient centrifugation following the manufacturer's protocol and cryopreserved in 10% DMSO. Cryopreserved PBMCs were thawed rapidly and then incubated for 30 minutes at 4 degrees with appropriate antibodies (**Table 3.1**). Fresh breast tumour biopsies were obtained and processed within 1 hour of surgical resection. Tumour tissue was first manually dissected and then chemically dissociated with a cocktail of Liberase DL/TL (Roche) for 30 minutes and then immunostained with anti-CD45 (**Table 3.1**) antibody. Dissociated cells were sorted for live singlet cells with appropriate gating strategy on BD FACSAria II (**Figure 3.1**).

#### Single cell sequencing

Single cell sequencing was performed by Dr Victor González-Huici. Sorted cells were counted on Countess and the gated population and the number of cells per sample that were loaded on Chromium 10X (version 2) were as follows: 1) HD4/5 (*CD45*+*CD3*+pan-γδ+; 10K cells from HD4 + 10K cells from HD5) and loaded on one lane; 2) HD6 (*CD45*+*CD3*+pan-γδ+; 20K cells were loaded on one lane); 3) BC1 (*CD45*+; 18K cells were loaded on one lane); 4) BC2 (*CD45*+; 12K cells

**Table 3.1:** Antibodies used in isolation of pan-gamma delta T cells from blood and validation.

| Antibodies | Clone | Cat# Biolegend | Fluorophore | Laser/filter | Dilution |
|---|---|---|---|---|---|
| anti-CD8 | SK1 | 344713 | APC-Cy7 | 640-780/60 | 1:200 |
| anti-GPR56 | 4C3 | 391905 | APC | 640-670/30 | 1:100 |
| anti-Vdelta2 | B6 | 331423 | PerCp-Cy5.5 | 488-695/40 | 1:200 |
| anti-CD16 | 3G8 | 302053 | PE-Dazzle 594 | 561-610/20 | 1:200 |
| anti-CXCR6 | K041E5 | 356013 | BV421 | 405-450/50 | 1:10 |
| anti-CD3 | OKT3 | 317339 | AF700 | 640-730/45 | 1:200 |
| anti-CD45 | HI30 | 304024 | AF700 | 619-710/50 | 1:10 |
| anti-TCRgd | REA591 | 130-113-512 Miltenyi Biotech | PE | 561-670/14 | 1:10 |

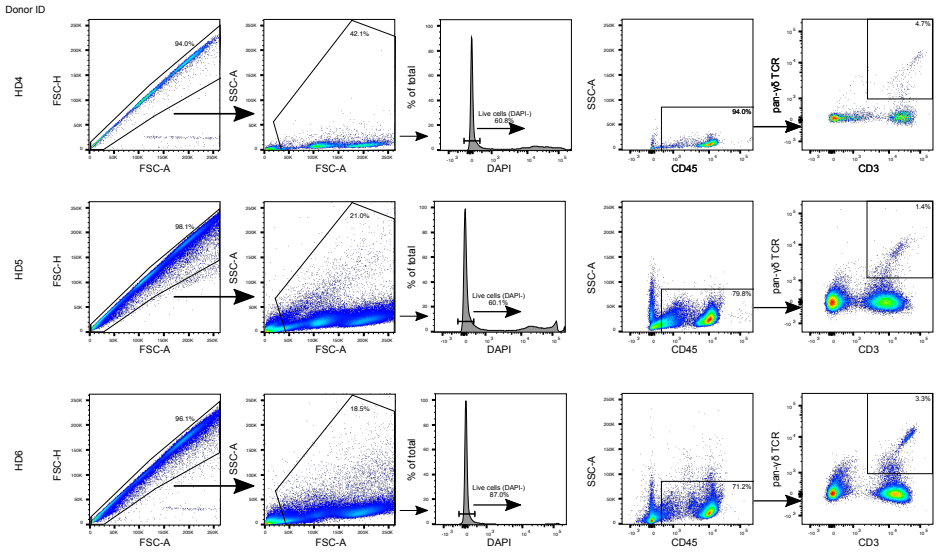



**Figure 3.1:** FACS gating strategy shown for the three PBMC donor samples, HD4, HD5 and HD6, that were subjected to 10X based single cell RNA-sequencing.

were loaded on one lane). scRNA library generation was carried out according to the manufacturers guidelines. Libraries from HD4/5 and HD6 were sequenced on NovaSeq S1 with 26bp + 90bp reads and an average of 42k reads per cell for HD4/5 and 74k reads per cell for HD6. Libraries from BC1 and BC2 were

sequenced on HiSeq2500 with 75bp + 75bp reads and an average of 53k reads per cell for BC1 and 47k reads per cell for BC2.

The number of cells loaded on Chromium 10X and the sequencing depth were decided after consultation with 10X Genomics representatives. Sequencing saturation ranged between 88.3% and 91.2% in the four different data sets. Since the multiplet rate increases with increased number of cells loaded, a QC step with stringent filtering thresholds, as described in the next section, was required to eliminate the number of multiplets included in downstream analyses.

Raw FASTQ format sequence reads were processed using the Cell Ranger (Zheng et al., 2017) pipeline for human genome hg19 assembly.

**Flow cytometry based validation of clusters**

Top differentially expressed surface markers for which antibodies are commercially available were selected and validated by Marcus Lindberg. Whole blood samples were collected with EDTA from 3 healthy donors. PBMCs were purified from whole blood using Ficoll-Paque PLUS. PBMCs were washed with 1x Dulbeccos PBS and cryopreserved in 10% DMSO at -80°C.

Fluorophores were optimised for the panel using the *BioLegend Fluorescence Spectra Analyzer* (BioLegend, 2019. http://www.biolegend.com/spectraanalyzer) to minimise spectral overlap. Antibodies were titrated on healthy donor PBMCs. PBMCs were thawed and rested for 30 minutes in fresh DMEM supplemented with 10% FBS. Cells were filtered with a 40µm cell strainer and incubated with *Human TruStain FcX* (BioLegend) for 20 minutes prior to staining. Cells were incubated with antibody (**Table 3.1**) for 30 minutes at 4°C and washed twice with FACS buffer (1X PBS, 1% BSA, 1 mM EDTA) prior to a final resuspension in FACS buffer. Compensation was performed using *UltraComp eBeads* (Invitrogen) compensation beads incubated with 1µL of antibody using the same protocol as

was used with the PBMCs. Samples were analysed on a *BD LSRFortessa* cell analyser. Flow cytometry files (.fcs) were analysed using *FlowJo* (version 10.0).

#### Data availability

Raw sequence data in `FASTQ` format and gene counts in matrix market format produced by the Cell Ranger pipeline (for hg19 assembly) have been deposited in GEO with study accession number GSE141665. 1) HD45 (GSM4210786) 2) HD6 (GSM4210787), 3) BC1 (GSM4210788) and 4) BC2 (GSM4210789).

### 3.2.2 Computational Analysis

#### Quality control (QC)

Counts data from the Cell Ranger pipeline were further filtered based on the number of genes and housekeeping genes as well as the percentage of mitochondrial genes. The list of human housekeeping genes was obtained from Tirosh et al. (2016). The list of mitochondrial genes was obtained from each dataset by selecting all genes whose names start with "MT". Different thresholds were selected for each dataset to account for the individual technical characteristics, based on manual inspection of the distributions of the number of genes and housekeeping genes and the proporion of mitochondrial genes per cell. **Figure 3.2** shows the thresholds selected for each dataset. Dots indicate cells and the retained cells are those inside the red box.

From the analysis of the remaining cells, I did not observe any clusters expressing a mixture of signatures of other clusters, which would indicate presence of doublets. I thus did not deem necessary to apply any other method for doublet detection.

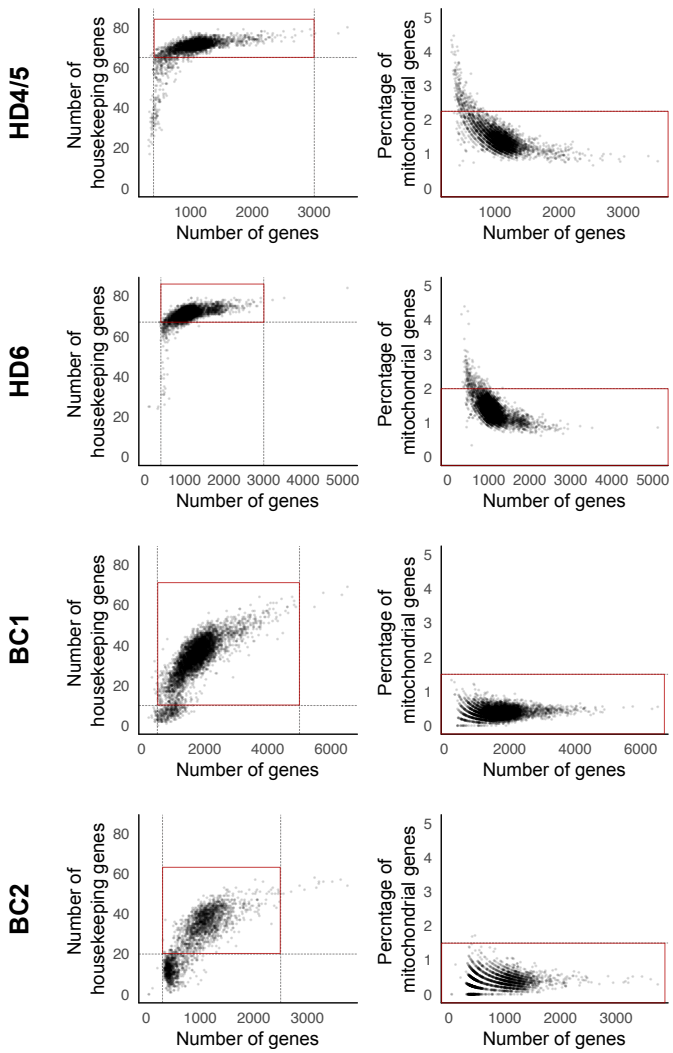



**Figure 3.2:** Quality control of all datasets used in this chapter. Cells were filtered based on the number of genes, the number of housekeeping genes and the percentage of mitochondrial genes. Different thresholds were selected for each dataset, indicated by the grey dotted lines. 3533 cells from HD4/5, 6147 cells from HD6, 4608 cells from BC1 and 1257 cells from BC2 that fall in the red boxes are those that were retained for further analysis.

### Annotation of cells positive for the various TCR gamma and delta gene segments

To determine the identity of the TRD and TRG gene segments present in each cell, `FASTQ` reads from each cell were aligned to the VDJ gene sequences (refdata-cellranger-vdj-GRCh38-alts-ensembl-2.0.0 from Cell Ranger) using `BWA`. Based on

the *CD45+* clusters in the BC1 data, only the *TRDV1*, *TRDV2*, *TRGV4* and *TRGV9* gene sequences were specifically enriched in the *TRDC+* clusters. Thus only reads uniquely mapping (mapping quality $> 30$) to *TRDV1*, *TRDV2*, *TRGV4* and *TRGV9* gene sequences were retained for further analysis in all four datasets.

### Alignment and clustering of scRNA-seq datasets

The two PBMC γδ-T cell scRNA-seq datasets (HD4/5 and HD6) were obtained from the same lab and with the same experimental protocol. Additionally, the cell composition is expected to be similar and the numbers of cells are relatively balanced. Thus, I chose to use alignment to remove the batch effect between the two datasets, which will help distinguish small and transcriptionally similar cell types due to the increased number of cells.

I used the Canonical Correlation Analysis (CCA, Butler et al. (2018)) from the Seurat package (version 3.1.0) to align and cluster the two datasets combined. Default parameters and 20 principal components were used. *TRDC* positive clusters from the merged dataset were retained. The final PBMC γδ-T cell data used here consisted of 6116 cells.

Similarly, the two breast tumour scRNA-seq datasets (BC1 and BC2), with a total of 6677 cells, were integrated and clustered using CCA from the Seurat package using the first 20 principal components that were significantly contributing to the explained variance of the data (Jackstraw plot method of Seurat).

### Identification and functional annotation of cluster-specific gene sets

The `MAST` function implemented in the Seurat package (version 3.1.0) was used to identify genes differentially upregulated between δ1 and δ2 subtypes. For the breast tumour data, depending on the required analysis, γδ-T cell clusters were either compared to each other in order to obtain gene sets that can

distinguish between γδ-T cell subtypes (conditional to the pan-gamma-delta markers) or to all immune cells in the datasets. More specifically, due to the high transcriptional similarity between γδ-T cell clusters and their relatively low abundance compared to other immune cell populations in the breast tumour data, I identified differentially expressed genes only between these three γδ-T cell clusters to identify potential markers for isolation and subtype-specific functions. On the other hand, for survival analysis, I used gene sets that distinguish each γδ-T cell cluster from all other immune cell populations in our data, to avoid detecting signals of other populations in TCGA samples (Ciriello et al., 2015) as will be explained in the Survival Analysis section of the Methods.

DAVID (Huang et al., 2009) was used to obtain enriched Gene Ontology (GO) Biological Process (GO-BP) and KEGG pathway terms for each extracted geneset. Given the list of genes differentially upregulated in a cluster as target and the list of all genes observed in the respective dataset as background, a list of GO-BP terms (GOTERM_BP_FAT) and KEGG pathway terms that were significantly enriched in the target list were downloaded as a functional annotation chart. Only terms with Benjamini adjusted p-values $\leq 0.05$ were retained.

**Enrichment of selected functional gene sets in the γδ-T cell subtypes**

scID was used to calculate an enrichment score per cell for *IFNγ* production, *IL17A* production, cytotoxic, adaptive and innate gene sets. Since there is no prioritisation of the contribution of each gene in the pathway, equal weights (equal to 1) were used for all the genes of each set. The genes used in each gene set are shown in **Table 3.2**.

**Table 3.2:** List of genes selected for each functional gene set.

| **Geneset** | **Genes** |
|---|---|
| *IFNγ* production | *TBX21*, *EOMES*, *STAT1*, *STAT4*, *IL12RB*, *IFNG* |
| *IL17A* production | *RORC*, *IL23R*, *CCR6*, *IL1R1*, *RORA*, *BLK*, *IL17A* |
| cytotoxicity | *GZMA*, *GZMB*, *GZMK*, *GZMM*, *GZMH*, *PRF1*, *TRAIL*, *FAS*, *IL12* |
| antigen presentation | *HLA-DQPB1*, *HLA-DRA*, *HLA-DPA1* |
| innate | *KLRK1*, *HCST* (*DAP10*) |

### Enrichment of published *CD28+* and *CD16+* δ2 subtype gene sets in the PBMC γδ-T cell subtypes

scID was used to calculate an enrichment score per cell for the *CD28+* and *CD16+* δ2 subtype gene sets from the study of Ryan et al (Ryan et al., 2016). Positive markers for the two subtypes shown in **Table 3.3** were used as input to scID without a reference single-cell RNA-sequencing data set. Weights were inferred from the target (PBMC) dataset.

**Table 3.3:** List of published *CD28+* and *CD16+* δ2 subtype gene sets from Ryan et al (Ryan et al., 2016).

| **Gene set** | **Genes** |
|---|---|
| *CD28+* | *GZMK*, *CD27*, *LTB*, *CCR7*, *MYC*, *CCR6*, *CD160*, *CD7*, *IL12RB2*, *CD28*, *CXCR6*, *RORC*, *SIPA1*, *IL7R*, *IL23R*, *IRF1*, *IL18R1*, *CCR2* |
| *CD28+* | *CX3CR1*, *GZMB*, *KIR3DL1*, *KIR2DL4*, *GNLY*, *KIR2DL3*, *KIR2DL1*, *KIR3DL2*, *GZMH*, *KIR2DS5*, *LAIR2*, *KIR3DL3*, *ITGB1*, *PRF1*, *SMAD7*, *BCL9L*, *LY6E*, *IL18RB*, *FCAR*, *KIR2DL5A*, *ITGAM*, *CCL4L2* |

### Survival analysis

Survival analysis was performed to test if presence or absence of any breast tumour γδ-T cell subtype was associated with survival rates of patients with breast cancer. RNA-seq and clinical data for TCGA breast

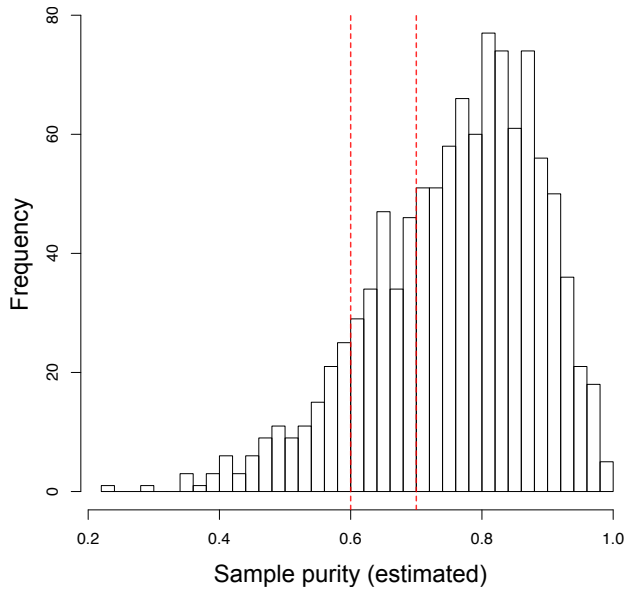



**Figure 3.3:** Histogram of tumour purity ($x$-axis) of TCGA breast tumour samples (Ciriello et al., 2015). A total of 191 samples with purity between 0.6 and 0.7, indicated by the red lines, were retained for further analysis.

cancer study (Ciriello et al., 2015) were obtained from NCI GDC Data Portal (https://portal.gdc.cancer.gov/projects/TCGA-Breast Cancer (BRCA)). Samples were filtered based on purity with the following criteria. Samples with low purity were removed as they are unlikely to reflect immune cells altered by presence of tumour. Additionally, samples with very high purity were removed as they are expected to contain only a small proportion of immune cells that is unlikely to consist of γδ-T cells given their very low abundance. This filtering resulted in 191 samples with purity between 0.6 and 0.7 (**Figure 3.3**).

I next calculated an enrichment score for each of the retained samples and each of the three γδ-T subtype gene signatures, as the weighted average of the expression of these gene signatures in each TCGA sample, with weights being the $log_e$ fold change values obtained from differential expression analysis of the scRNA-seq data. Then, for each signature, I selected the bottom and top 1/3 of the samples

to represent the samples with low and high expression of the gene signature respectively and performed survival analysis between the two groups using the `survfit` function of the `survival` R library (Version 2.44.1.1).

### Mapping across datasets with scID

To identify BC-equivalent γδ-T subpopulations in the PBMC I used scID and calculated a score of each PBMC cell for each of the BC γδ-T subcluster signature. Each signature consists of genes that are upregulated and genes that are downregulated in the reference BC cluster, thus positive scID scores indicate cells that express more upregulated than downregulated genes, and negative scores indicate cells that express more downregulated than upregulated genes.

### Prediction of signalling between γδ-T cells and other immune cells

CellPhoneDB (Efremova et al., 2019) was used to calculate the likelihood of cell type specificity of receptor-ligand interactions between γδ-T and other tumour infiltrating immune cell types. Gene expression data and cell type labels for the BC1 and BC2 cells were used as input to CellPhoneDB `statistical_analysis` method. Only interactions with p-values $\leq 0.01$ were retained.

## 3.3 Results

### 3.3.1 Identification of subtypes of γδ-T cells in peripheral blood via unsupervised clustering of scRNA-seq data

Peripheral blood γδ-T cells from 3 healthy donors were sorted using a pan-gamma-delta antibody (**Table 3.1**) and subjected to droplet-based scRNA-seq using 10X Chromium (Zheng et al., 2017) single cell 3' gene expression library generation and subsequently sequenced on the Illumina Nova-Seq S1 platform. Cells from donors HD4 and HD5 were pooled before performing scRNA-seq and labelled as HD4/5. Raw sequence reads were processed using the Cell Ranger pipeline (see Methods). The resulting gene expression counts for each dataset were merged and clustered using Seurat CCA (Butler et al. (2018), Stuart et al. (2019)). Only clusters that were *TRDC* positive and present in all datasets were retained for subsequent analysis. This final analysis-ready dataset comprised of 6,116 cells with an average of 1,084 genes expressed per cell representing 12,943 different genes in total.

Two-dimensional projection of the merged scRNA-seq data using UMAP (Becht et al., 2019) revealed two macroclusters (**Figure 3.4 A**) containing 5 clusters in total. Cells from each of the 2 datasets (HD4/5 and HD6) were relatively evenly distributed among the clusters and thus biologically reproducible (**Figure 3.4 B**). All clusters had relatively uniform distribution of TCR delta constant gene segment (*TRDC*) and therefore are bonafide γδ-T cells. A macrocluster positive for TCR δ2 variable gene segment (*TRDV2*) and a small macrocluster positive for TCR δ1 variable gene segment (*TRDV1*) could be identified (**Figure 3.4 C**). Cells expressing genes that encode the other delta chains could not be unambiguously identified in this short read (90bp) sequencing data. Thus we

observe three *TRDV2* positive (here on referred to as δ2) clusters and two *TRDV1* positive (here on referred to as δ1) clusters.

As the sequencing coverage of individual cell transcriptomes are highly sparse and the expression levels of the TRD and TRG genes may be different and thus have different dropout rates, I determined their presence independently at the cluster level. Only the *TRGV4* and *TRGV9* gene segments could be unambiguously aligned with the 3' UTR short read sequences that comprised this data (see Methods). Computation of the proportion of cells in each cluster positive for the various TRG gene segments (**Figure 3.4 D**), revealed that all of the three δ2 clusters were preferentially enriched in *TRGV9* positive cells consistent with the reports in literature that the majority of PBMC γδ-T cells are *TRDV2-TRGV9* positive (Dimova et al., 2015). In contrast, the δ1 clusters were enriched for *TRGV4* also consistent with the reported preferential pairing of Vδ1 and Vγ4 chains (Jiang et al., 2012). The differential enrichment of the TRGV gene segments between the δ1 and the δ2 was statistically significant (Chi-square test for independence, $\chi 2=17.836$, P=0.000134, dof=2). Thus, we observe three Vδ2Vγ9 and two Vδ1Vγ4 clusters in PBMC.

Differential gene expression analysis uncovered several cluster specific enriched genes (**Figure 3.5 A**), which could be potentially used as markers for isolation of these subsets for further functional validation. The complete list of differentially expressed genes between the PBMC γδ-T cell subtypes can be found in **Table A.1**. Functional annotation of these gene sets suggests that these five γδ-T cell subtypes have different functions (**Figure 3.5 B, Table A.6**).

While the δ2.1 cluster has a distinct gene expression pattern that could be either due to its abundance or due to being transcriptionally very different from the other clusters, there is high transcriptional similarity between δ1.1 and δ1.2 and between δ2.2 and δ2.3 clusters indicated by the large number of shared markers

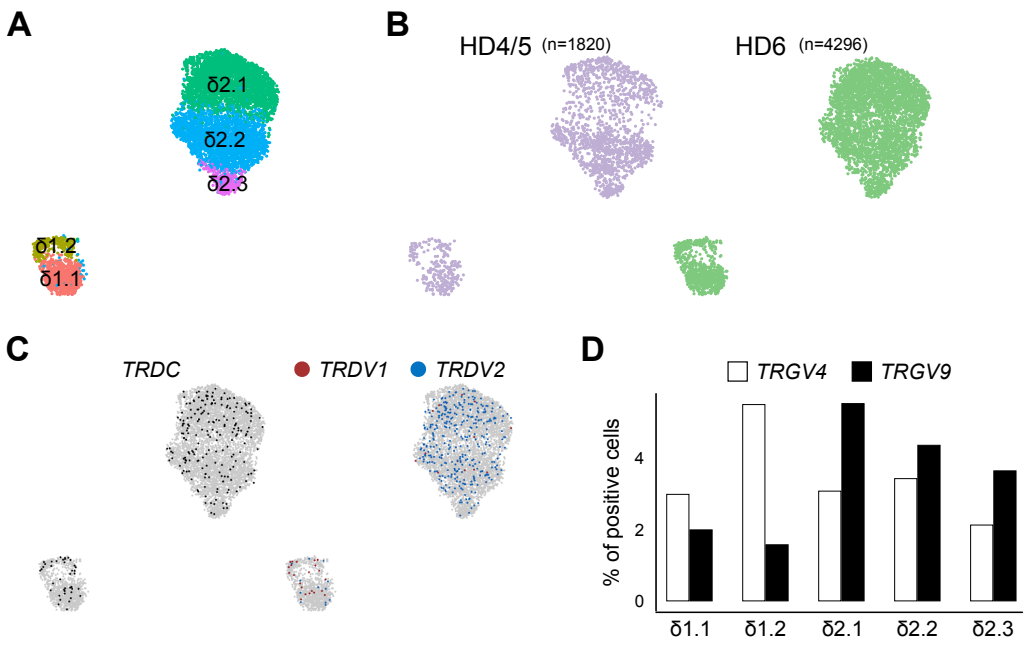



**Figure 3.4:** Unsupervised analysis of scRNA-seq data on γδ-T cells from peripheral blood of healthy adult donors identifies multiple δ1 and δ2 subtypes. **(A)** UMAP of the merged single cell gene expression data of PBMC derived γδ-T cells from 3 healthy donors. Different clusters are named according to TCR delta chain identity. *TRDV1* (gene that encodes δ1 chain) positive clusters were labelled with prefix δ1 and *TRDV2* (gene that encodes δ2 chain) positive clusters were labelled with prefix δ2. **(B)** Overlay of data source on UMAP of scRNA-seq data. Cells from donor HD4 and HD5 were pooled before performing scRNA-seq and are labelled as HD4/5. The number of cells from each dataset is shown above the projection. **(C)** Identification of cells positive for TCR delta genes. Overlay of cells that have genes mapping to the *TRDC* (left) and *TRDV1/2* (right) gene segments. Grey indicates absence and colour indicates presence of reads mapping to *TRDC* (black), *TRDV1* (red) and *TRDV1/2* (blue) gene segments. **(D)** Quantification of enrichment of genes expressing TCR gamma (γ) chain in each cluster. Only the TCR gamma genes that could be unambiguously mapped (see Methods) were considered. *Y*-axis shows the percentage of cells within each cluster (*x*-axis) of the merged data that is positive for *TRGV4* (white) or *TRGV9* (black) gene segments.

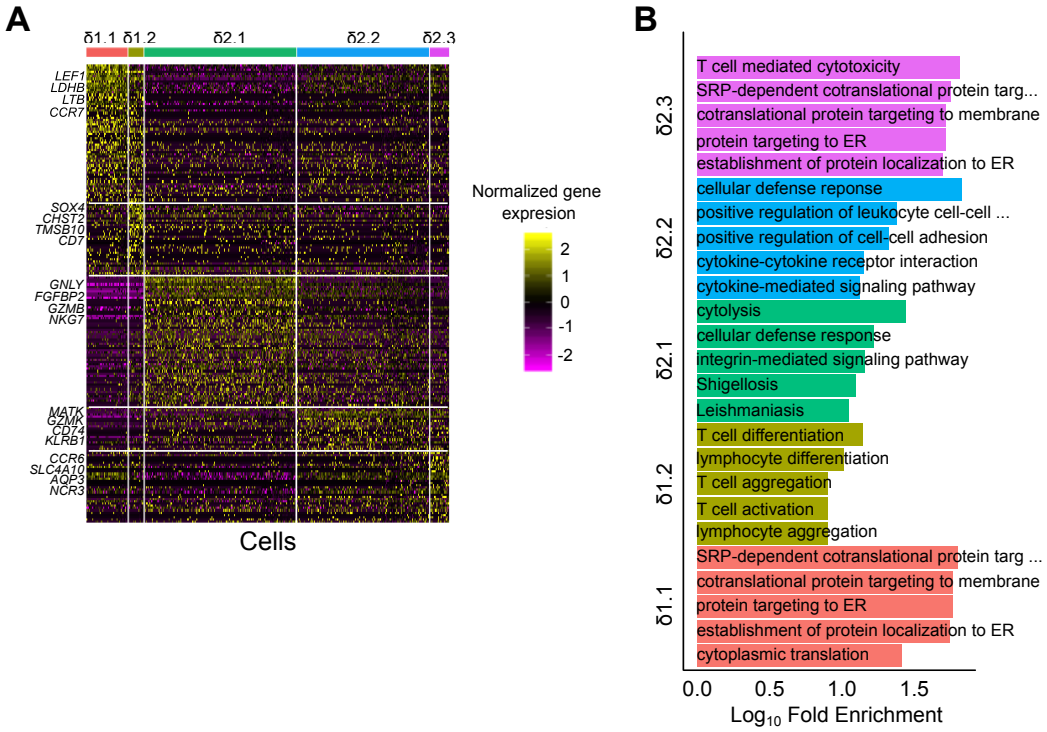



**Figure 3.5:** **(A)** Heatmap showing genes (rows) enriched in each of the clusters (columns). Yellow represents enrichment and purple represents depletion. Top 4 genes per cluster are labelled. **(B)** Functional annotation, as defined in GO and KEGG databases, of cluster specific differentially expressed genes. The top 5 significantly enriched functional terms are shown.

(**Figure 3.5 A**). In order to find markers and functions that better distinguish between these similar subtypes, I also obtained differentially expressed genes for these pairs of clusters separately and performed functional annotation of these gene sets. **Figure 3.6** shows the differentially expressed genes and annotated functions between δ1.1 and δ1.2 (see also **Table A.2 and Table A.7**) and **Figure 3.7** shows the differentially expressed genes and annotated functions between δ2.2 and δ2.3 (see also **Table A.3 and Table A.8**).

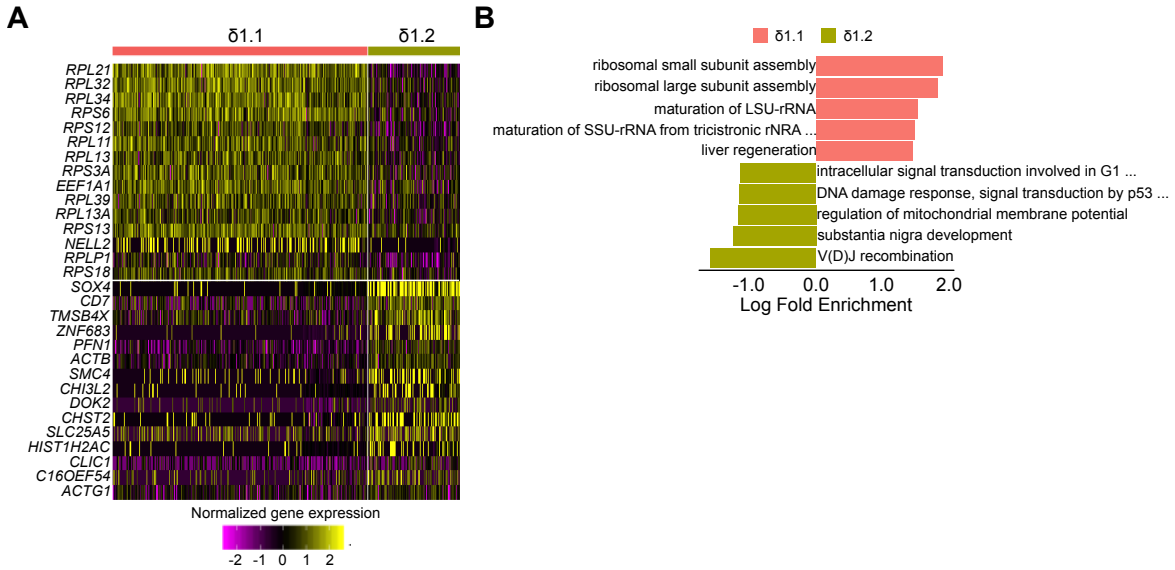



**Figure 3.6:** Comparison of δ1.1 and δ1.2 clusters. **(A)** Heatmap showing top 15 genes (rows) differentially expressed between the two subtypes (columns). Yellow represents enrichment and purple represents depletion. **(B)** Functional annotation of genes differentially expressed between the two subtypes. The lengths of the bars ($x$-axis) are proportional to the log fold enrichment. The negative values indicate enrichment in δ1.2 (khaki) and the positive values indicate enrichment in δ1.1 (red) cluster. Terms are as defined by the Gene Ontology and KEGG databases.

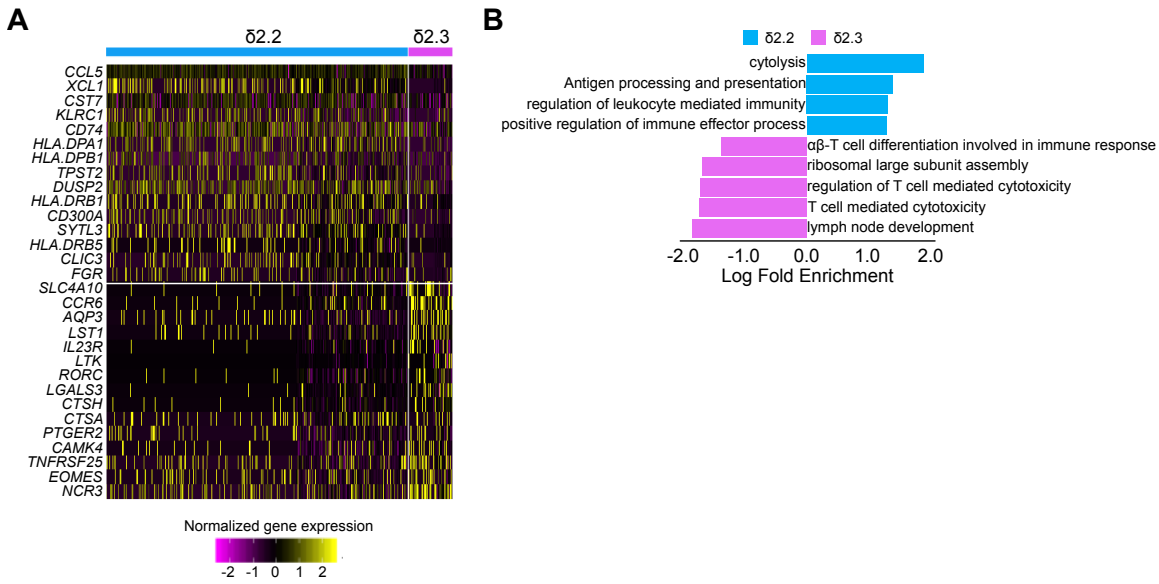



**Figure 3.7:** Comparison of δ2.2 and δ2.3 clusters. **(A)** Heatmap showing top 15 genes (rows) differentially expressed between the two subtypes (columns). Yellow represents enrichment and purple represents depletion. **(B)** Functional annotation of genes differentially expressed between the two subtypes. The lengths of the bars ($x$-axis) are proportional to the log fold enrichment. The negative values indicate enrichment in δ2.3 (purple) and the positive values indicate enrichment in δ2.2 (blue) cluster. Terms are as defined by the Gene Ontology and KEGG databases.

Finally, I used curated gene sets for *IFNγ* production, *IL17A* production, cytotoxicity, antigen presentation and innate pathways to annotate the clusters based on the enrichment of these gene sets in the cells within each cluster (**Figure 3.8**).

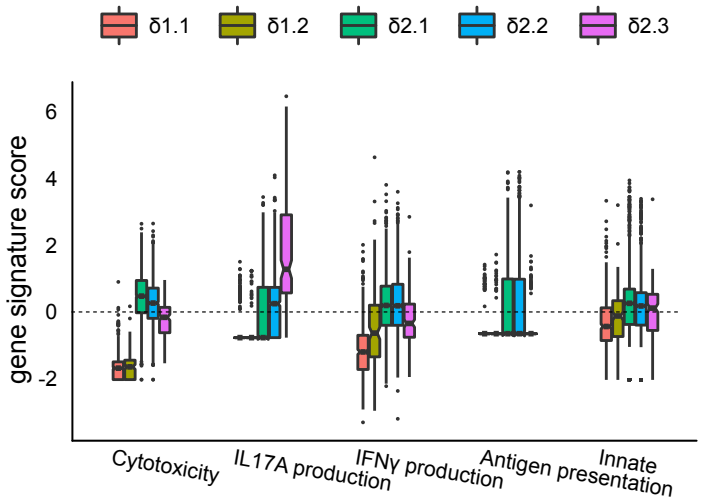



**Figure 3.8:** Distribution of gene signature scores ($y$-axis) for Cytotoxicity, IL17A production, IFNγ production, Antigen presentation on MHC class 1 and Innate gene sets (**Table 3.2**) in each of the PBMC γδ-T cell cluster ($x$-axis). scID but with equal weights was used to calculate an enrichment score per cell for each of the gene signatures (see Methods)

.

All the above results suggest the following annotation: i) the δ1.1 cluster appears to be naive or immature, indicated by expression of *LEF1* and *CCR7*, ii) the δ1.2 cluster could be recent thymus emigrant (expression of *CD7*) , iii) the δ2.1 cluster appears to carry out canonical δ2 function, namely anti-bacterial and antigen presentation function, and interferon gamma (IFNγ) production, and is the most cytotoxic of all the other PBMC γδ-T cell clusters, iv) the δ2.2 cluster appears to be involved in antigen presentation and, v) the δ2.3 cluster is a clear IL17A producer. This IL17A producing subtype is also differentially enriched in *CCR6* and consistent with a previous report (Haas et al., 2009).

The above clustering was validated using FACS by selecting surface proteins from the list of differentially expressed genes for which antibodies are commercially available. More specifically antibodies against *CXCR6* (present in both δ2.2 and δ2.3) and *GPR56* (present in δ2.1) were selected (**Figure 3.9 A**). PBMCs from three donors (including two additional healthy donors, HD9 and HD10) were stained with anti-CD3, anti-TCRγδ-δ2, anti-GPR56, and anti-CXCR6 (**Figure 3.9 B**). The L-shaped scatter plot of *GPR56* and *CXCR6* suggests mutual exclusion and corroborates that they mark different δ2 sub-populations within the human blood γδ-T population.

I additionally used literature to further validate the clusters identified in this scRNA-seq data. The IL17A producing cluster (δ2.3) differentially expressed *CCR6* as reported previously (Haas et al., 2009). Ryan et al (Ryan et al., 2016) identified two different δ2 subtypes in PBMC γδ-T cells that are marked by mutually exclusive expression of *CD28* and *CD16*, which were also observed in the data presented in this chapter (**Figure 3.9 C**). While *CD16* was specifically enriched in only one δ2 subtype (i.e. δ2.1), *CD28* was more diffused and also present in the δ1 subtypes. To corroborate this with a larger gene set, I scored all the subclusters for the published gene signatures of the γδ-T-cell partitioning defined by *CD16* and *CD28* (Ryan et al., 2016). While the *CD16* gene signature was exclusive to one subtype of γδ-T cells, the *CD28* gene signature was enriched in all the other subtypes, including the two δ1 subtypes (**Figure 3.9 D**), hence *CXCR6* is better than *CD28* as a marker for isolating *CD16* negative δ2 cells, because unlike *CD28*, its expression is absent in δ1 γδ-T cells.

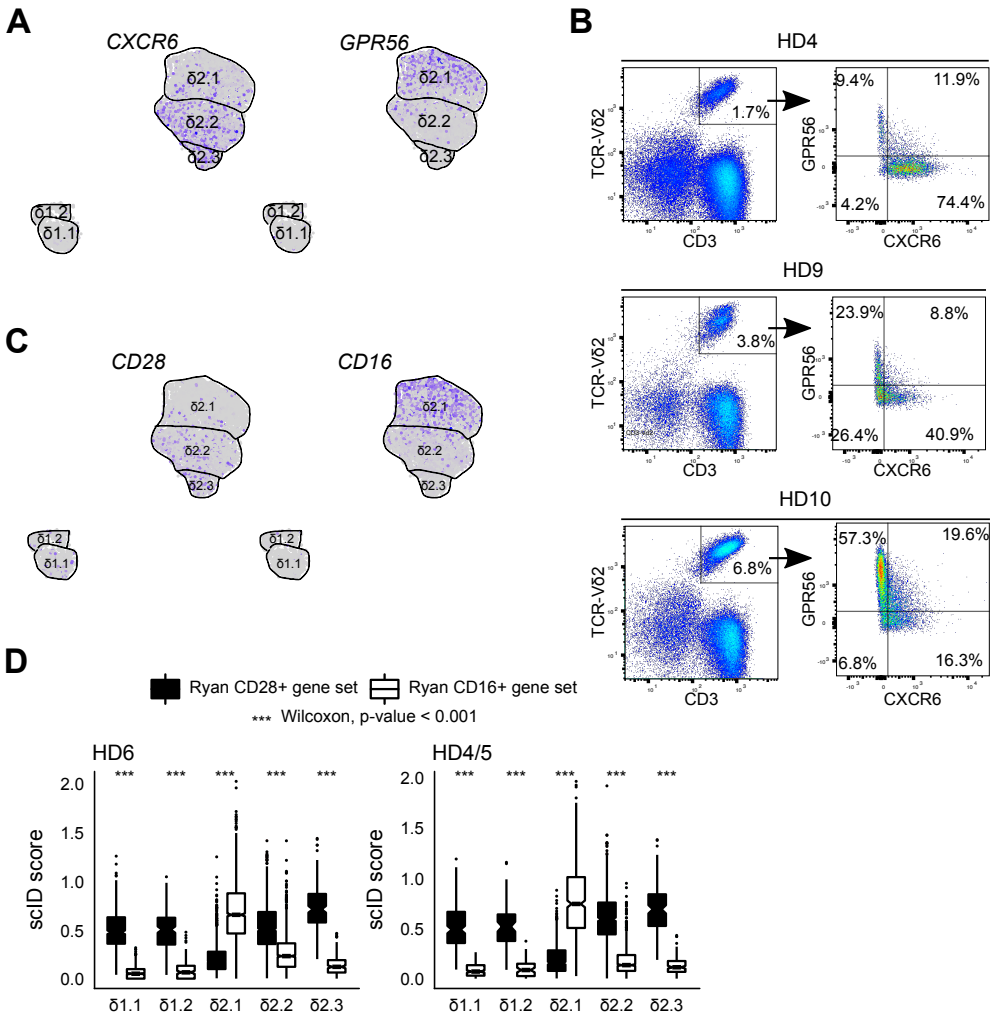



**Figure 3.9:** Validation of the PBMC γδ-T subtypes. **(A)** Feature plot showing expression of *GPR56* and *CXCR6*, markers that appear to be mutually exclusive in PBMC δ2 subtypes. **(B)** Flow cytometry based validation of novel markers, *GPR56* ($y$-axis) and *CXCR6* ($x$-axis), of peripheral blood γδ-T δ2 subtypes. Healthy donor identities are indicated in the title. Numbers in each quadrant indicate percentage of δ2 cells. This work was performed by Marcus Lindberg. **(C)** Feature plot showing *CD16* and *CD28*, which are published markers of the δ2 subtype of PBMC γδ-T cells (Ryan et al., 2016)). **(D)** Scores for published gene signatures of CD16 (white) and CD28 (black) δ2 subtypes in the clusters found in our PBMC γδ-T scRNA-seq data ($x$-axis). P-values were computed using the Wilcoxon signed-rank test.

In summary, I have provided evidence that the classification suggested in this chapter is consistent with what has been reported in the literature regarding human PBMC γδ-T cells. Additionally, the genes differentially expressed in the two *CD28* positive δ2 subtypes of PBMC γδ-T cells further suggest that they are two functionally different populations as one subtype was *IFNγ+* while the other was an IL17A producer, observations that were missed from previous studies.

### 3.3.2 Identification of γδ-T cell subtypes within breast tumour microenvironment

To uncover γδ-T cell subtypes and their gene expression programs in the tumour microenvironment, we dissociated fresh Triple Negative Breast Cancer (TNBC) and Her2+ breast tumour biopsies into single cells and performed scRNA-seq on all immune cells (CD45+) (see Methods). Three clusters that were double positive for *CD3* and *TRDC* were present in both the samples and were designated as γδ-T cells (**Figure 3.10 A,B,C**). While I was able to detect reads mapping uniquely to *TRDV2*, I did not identify *TRDV1* transcripts despite being able to detect these in the PBMC γδ-T cells. There are two possible explanations for this; either the shorter reads (75bp compared to 90bp in PBMC) do not include this region, or the variable region in *TRDV1* gene in breast tumour γδ-T cells is much further away from the 3' UTR than it is in the PBMC γδ-T cells. As seen in PBMC γδ-T cells, the *TRDV2*+ cluster (i.e. γδ-T.3) is also *TRGV9*+, while the other two clusters (γδ-T.1 and γδ-T.2) are *TRGV4*+ (**Figure 3.10 D**). The preferential pairing of TCR δ2 and γ9 chains and of TCR δ1 and δ4 chains seen in PBMC γδ-T cells (**Figure 3.4 D**) suggests that γδ-T.1 and γδ-T.2 are possibly *TRDV1*+. Additionally, presence of reads mapping to TCR δ and γ chains further supports the evidence that these three clusters consist of γδ-T cells.

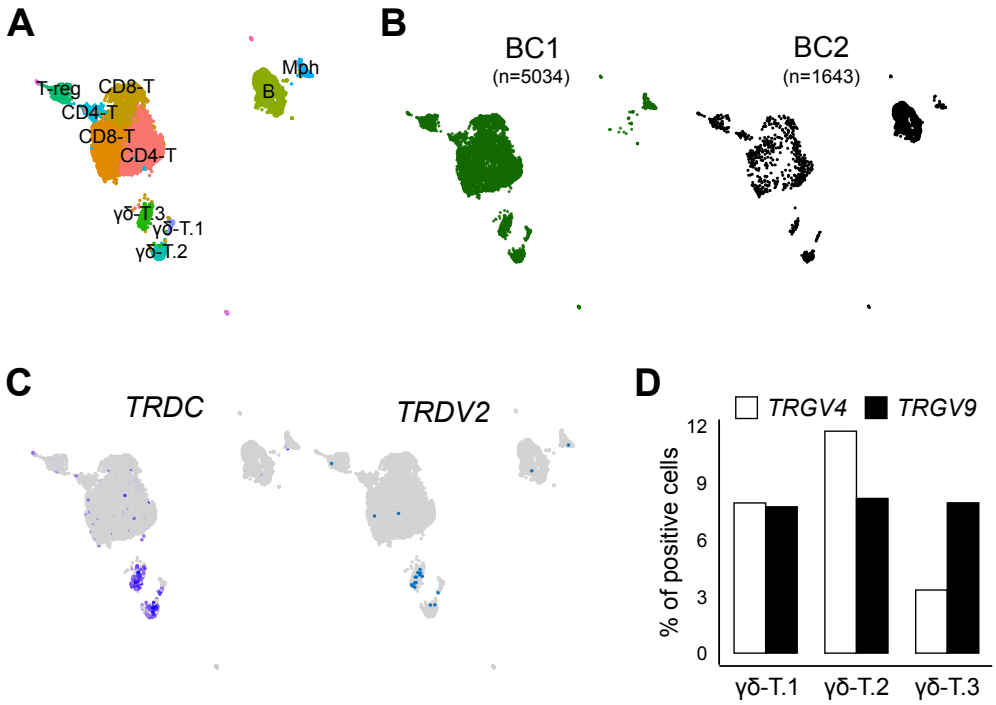



**Figure 3.10:** Unsupervised clustering of breast tumour infiltrating immune cells uncovers three subtypes of γδ-T cells. **(A)** UMAP of the merged single cell gene expression data of breast tumour infiltrating immune cell data sets from two patients. Three clusters were double positive for *CD3* and *TRDC* and were classified as γδ-T cells. Abbreviations: Mph, Macrophage, T-reg, regulatory T cells; B, B cells; CD8-T, *CD8*+ αβ-T cells; CD4-T, *CD4*+ αβ-T cells. **(B)** Overlay of donor identity on UMAP of scRNA-seq data. The number of cells from each donor is shown above the projection. BC1 is a triple negative subtype and BC2 is a Her2+ subtype of breast cancer (BC). **(C)** Identification of cells positive for genes encoding TCR δchain. Overlay of cells that have genes mapping to the *TRDC* (left) and *TRDV2* (right) gene segments. No *TRDV1* positive cells were identifiable. **(D)** Quantification of enrichment of genes encoding TCR γ chain in the BC γδ-T clusters. *Y*-axis shows the percentage of cells positive for the indicated *TRGV4* and *TRGV9* gene segment within each BC γδ-T cluster ($x$-axis).

Cluster specific differentially expressed genes (**Table A.4**) suggested that these clusters had different functional roles (**Figure 3.11 A,B**). The complete list of functional annotations significantly enriched in these gene sets can be found in **Table A.9**. γδ-T.1 had the highest levels of cytotoxicity. γδ-T.2 was the only innate-like subtype and had the highest levels of *IFNγ* production and antigen presentation. γδ-T.3 produced the highest level of *IL17A* relative to the other clusters (**Figure 3.11 C**).

Using mathematical deconvolution to infer the proportion of various immune cell types from bulk RNA-seq data, elevated levels of γδ-T cells were found to be associated with better survival in breast cancer (Ma et al., 2012) and across all cancers (Gentles et al., 2015); however, recent work observed an improved overall survival in triple-negative breast cancer for a δ1 innate-like subtype of γδ-T cells but not for the overall levels of γδ-T cells (Wu et al., 2019). I sought to repeat this analysis using the BC γδ-T subtype specific gene signatures identified in this chapter (**Table A.5**).

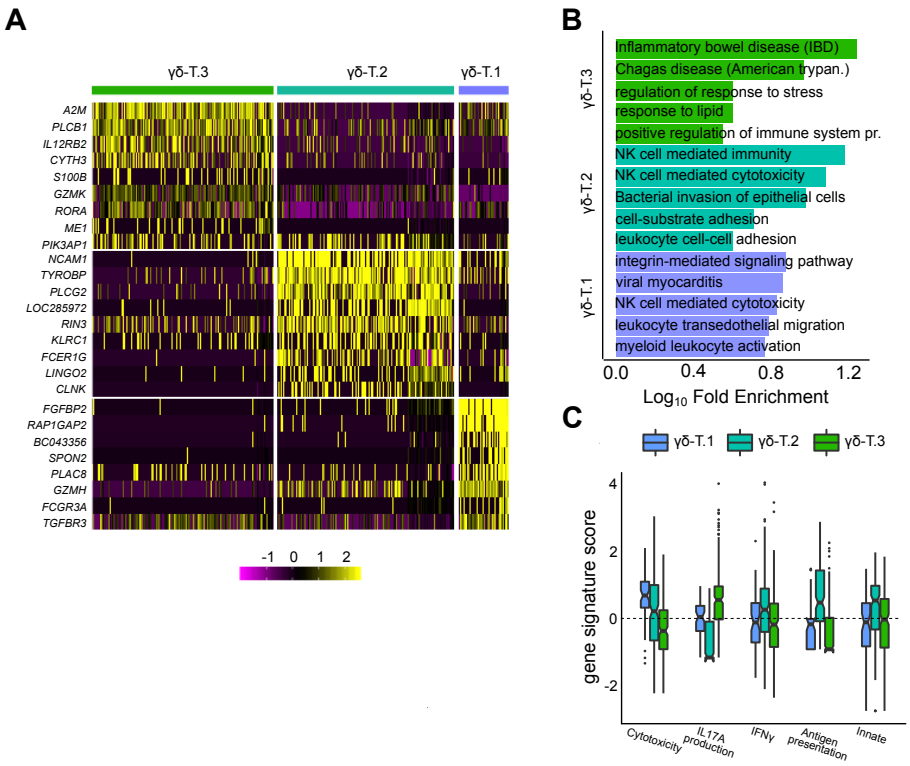



**Figure 3.11:** Identification of genes differentially expressed between the three breast tumour infiltrating γδ-T subtypes and functional annotation. **(A)** Heatmap showing top differentially expressed genes (row labels) between the three BC γδ-T cell subtypes. Yellow represents high expression and purple represents low expression. **(B)** Functional annotation of genes differentially expressed between the three BC γδ-T cell clusters using data from GO and KEGG databases. The lengths of the bars ($x$-axis) are proportional to the log fold enrichment. **(C)** Distribution of gene signature scores ($y$-axis) for *IFNγ* production, *IL17A* production, Cytotoxicity, Antigen presentation on MHC class 1 and Innate gene sets in each of the BC γδ-T cell cluster ($x$-axis).

Based on the enrichment score for each of the retained samples and each of the three BC γδ-T subtype gene signatures (see Methods) I selected the bottom and top 1/3 of the samples to represent the samples with low and high expression of the gene signature respectively (**Figure 3.12**) and performed survival analysis between the two groups.

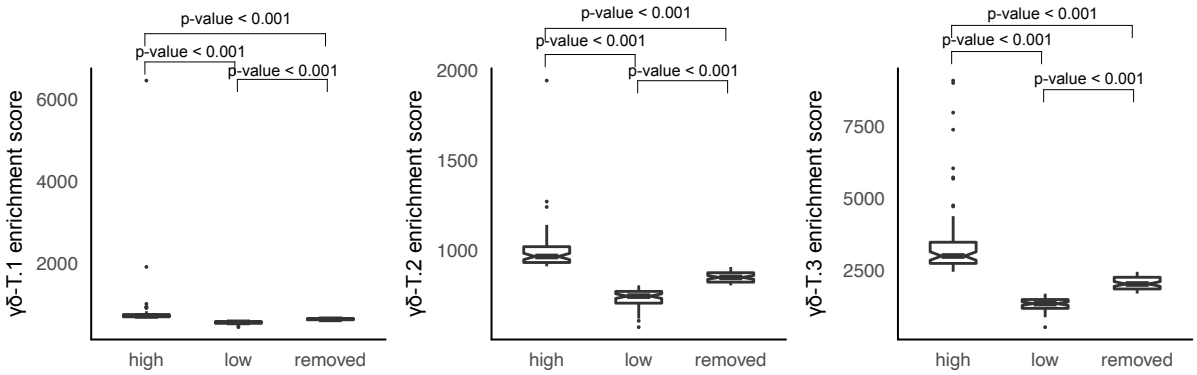



**Figure 3.12:** Enrichment scores of the retained TCGA breast cancer samples for each of the three breast tumour γδ-T subtype gene signatures. Wilcoxon rank test was used to compute statistical significance of the scores between the grouped samples.

Survival analysis of the breast cancer data from the Cancer Genome Atlas (TCGA) (Ciriello et al., 2015), indicated that only the enrichment of BC γδ-T.2 gene signature, but not the signature of either of the other two BC γδ-T clusters, was associated with improved survival (**Figure 3.13 A**). The group of patients with high expression of BC γδ-T.2 gene signature did not have higher overall levels of T-cells (**Figure 3.13 B**) or a higher mutation load (**Figure 3.13 C**) or higher expression of *NKG2D* ligands (**Figure 3.13 D**) than the group of patients with lower expression of BC γδ-T.2 gene signature, suggesting that the difference in survival rates observed is possibly not due to a secondary consequence of these previously well known factors associated with survival. These findings are also consistent with the findings reported in Wu et al (Wu et al., 2019).

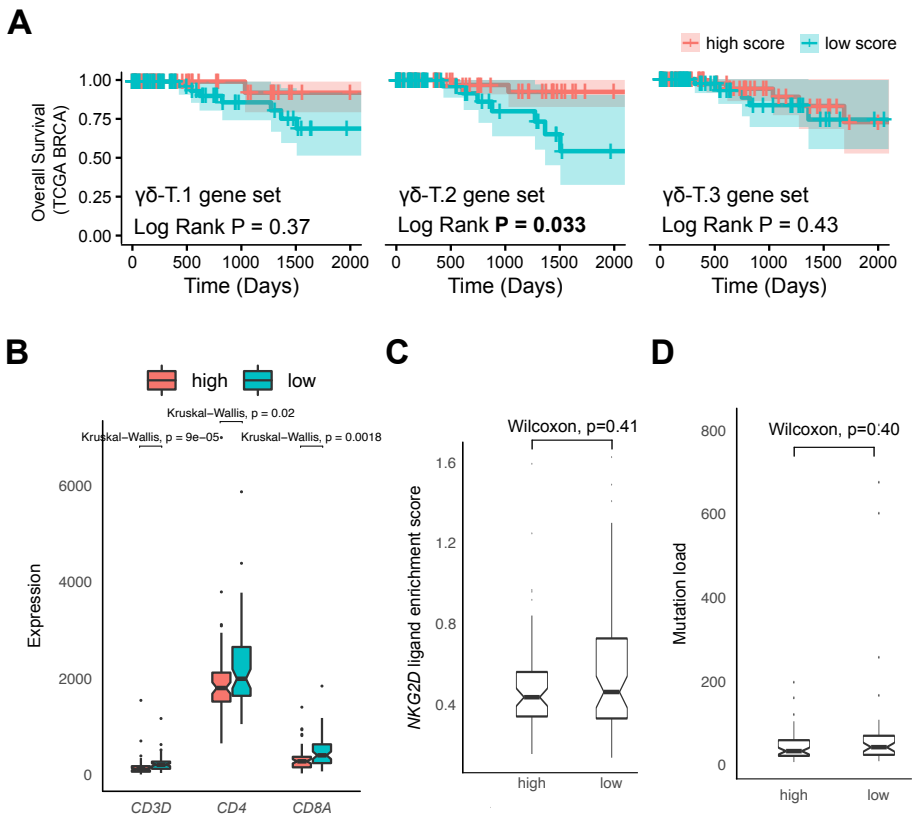



**Figure 3.13:** Characterisation of breast tumour infiltrating γδ-T cells uncovers a subtype of γδ-T cluster that is associated with favourable outcome. **(A)** Kaplan-Meir survival curve of the TCGA breast cancer data (Ciriello et al., 2015). Patients were partitioned into high and low group based on scores for gene signatures of each of the indicated BC γδ-T cluster. *Y*-axis shows overall survival. **(B)** Boxplot of expression (*y*-axis) of *CD3D*, *CD4* and *CD8A* genes in TCGA breast cancer data samples with high (red) and low (blue) expression of the BC γδ-T.2 subtype gene signature (*x*-axis). **(C)** Boxplot of scID scores (*y*-axis) of the *NKG2D* ligands gene set in TCGA breast cancer data samples with high and low expression of the BC γδ-T.2 subtype gene signature (*x*-axis). Wilcoxon rank test was used to compute statistical significance of different scores within each cluster. **(D)** Boxplot of mutation load (*y*-axis) of TCGA breast cancer data samples with high and low expression of the BC γδ-T.2 subtype gene signature (*x*-axis). Wilcoxon rank test was used to compute statistical significance of different scores within each cluster.

Sequencing unsorted breast tumour-infiltrating immune cells provided an additional opportunity to explore preferential ligand-receptor interactions between γδ-T and other immune cells using CellPhoneDB (Efremova et al., 2019). **Figure 3.14** shows the significant interactions ($p - value \leq 0.01$) between each γδ-T subtype and other immune cell subtypes in BC1 (left) and BC2 (right). Only ligand-receptor pairs that are unique to a specific γδ-T subtype are shown. Colour indicates the average expression of the ligand and the receptor in the interacting cell types and white represents non-significant interactions ($p-value > 0.01$). γδ-T.1 cells interact with various types of immune cells mainly through aMb2 ($\alpha_M\beta_2$) and aXb2 ($\alpha_X\beta_2$) complexes and the tumour necrosis factor receptor superfamily (TNF). aMb2 and aXb2 complexes are integrins expressed on the cell surface of various leukocytes, including macrophages and NK cells, playing a pivotal role in innate immunity, with the $\alpha_M$ and $\alpha_X$ subunits mediating adhesion and spreading of cells and the b2 subunit mediating cell migration (Solovjov et al. (2005), Hynes (2002)). γδ-T.2 cells interact mostly with macrophages and γδ-T.3 cells interact mostly through the *DPP4* receptor. However, there is very low confidence in the above observations as most of the interactions are restricted to only one of the BC samples.

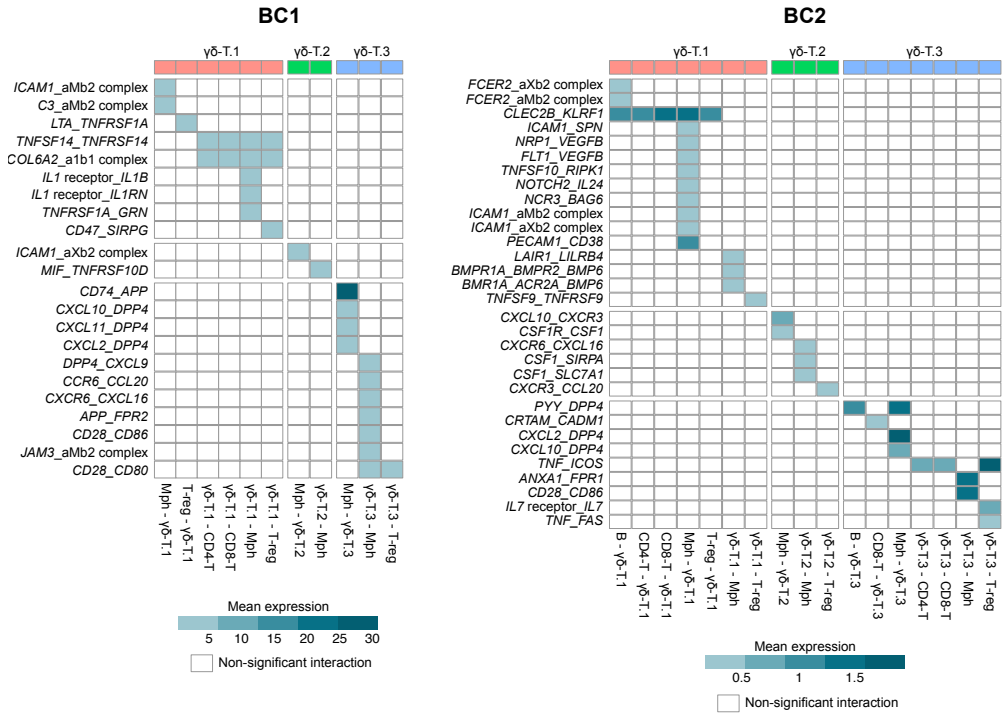



**Figure 3.14:** Identification of preferential ligand-receptor interactions between γδ-T and other immune cells using CellPhoneDB (Efremova et al., 2019). Heatmap showing average expression of the ligand and the receptor in the interacting immune cell types where the interaction is significant (blue scale). Only ligand-receptor pairs that are unique to a specific γδ-T subtype are shown. White represents non-significant interactions. Left panel shows interactions between immune cell types in BC1 and right panel shows interactions between immune cell types in BC2.

### 3.3.3 Comparison of PBMC and breast tumour γδ-T cells

Lastly, I sought to compare and contrast breast tissue infiltrating γδ-T cells with their counterparts in the circulation, starting with selected genes relevant for anti-tumour immune function. Overall, the BC γδ-T cells were more activated and had higher abundance of transcripts of genes involved in cytotoxicity and exhaustion, while the markers of memory and naive T cells were significantly lower in breast tumour γδ-T cells compared to those in PBMC (**Figure 3.15 A**).

To identify equivalent clusters across the BC and PBMC γδ-T datasets, I used scID to calculate matching scores of all PBMC cells from HD6 for the three BC γδ-T subtype gene signatures (**Figure 3.15 B**). I decided to use the BC subtypes as reference since the PBMC data represent a larger number of cells. Consequently, an inability to detect an equivalent BC subpopulation in the PBMC indicates absence. On the other hand, an inability to detect a PBMC subpopulation in the BC data could be due to sampling, i.e. no cells representing this cell type were selected due to low prevalence. Based on these scores, BC γδ-T.1 cluster seems to be transcriptionally equivalent to the PBMC δ2.1 cluster. This is supported by both clusters being marked by *CD16* and being IFNγ producers (**Figure 3.8**, **Figure 3.11 C**). The BC γδ-T.3 cluster seems equivalent to the PBMC δ2.3 cluster, consistent with both being marked by *CCR6* and being positive for *IL17A* (**Figure 3.8**, **Figure 3.11 C**). However, none of the PBMC clusters were similar to BC γδ-T.2, which is defined by expression of *NCAM1* (*CD56*).

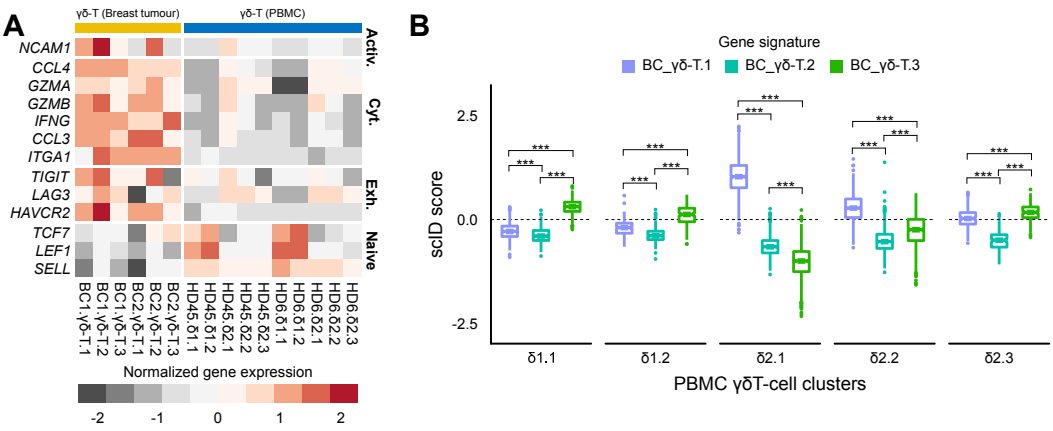
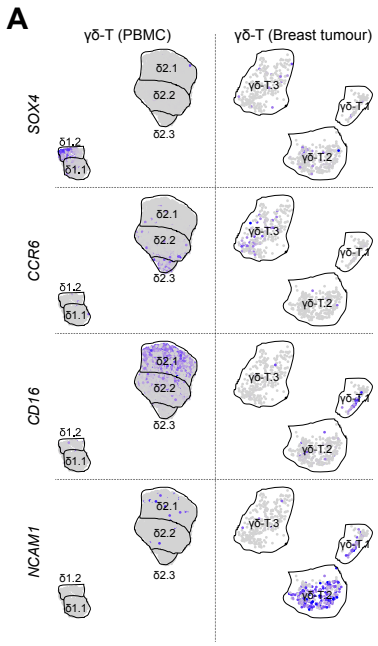



**Figure 3.15:** **(A)** Comparison of expression of genes (rows) involved in activation, cytotoxicity, exhaustion and naive T-cell state between PBMC γδ-T cell and breast tumour infiltrating γδ-T cell subtypes. Gray represents low average expression and red represents high average expression of the genes in each subtype (columns). **(B)** Assessment of similarity of the γδ-T cell subtypes in PBMC and breast tumour. Boxplot of scID scores ($y$-axis) of the BC cluster specific gene signatures in the PBMC clusters ($x$-axis). Scores above the dashed line indicates enrichment of the indicated gene signature. A Mann-Whitney U test was used to compute statistical significance of different scores within each cluster. "***" indicates P-value $\leq 0.001$.

All genes selected from this study that can help distinguish between subtypes of γδ-T in PBMCs and breast tumours are summarised in the feature plots in **Figure 3.16 A**. These results suggest a refined classification of human γδ-T cells based on the single-cell RNA-seq data presented in this chapter: *SOX4* and *NCAM1* can distinguish the three δ1 clusters, while the combination of *CD16* and *CCR6* can be used to define the three δ2 subtypes (**Figure 3.16 B**).




**B**

***Human γδ-T subtypes***

| **TCR-γδ** | *TRDV1 / TRGV4* | | | *TRDV2 / TRGV9* | | |
|---|---|---|---|---|---|---|
| **Marker genes** | *SOX4⁻ NCAM1⁺* | *SOX4⁺* | *NCAM1⁺* | *CD16⁺* | *CCR6⁻ CD16⁻* | *CCR6⁺* |
| **Tissue Presence** | Blood δ1.1 | Blood δ1.2 | Breast Tum BC_γδ-T.2 | Blood δ2.1 Breast Tum BC_γδ-T.1 | Blood δ2.2 | Blood δ2.3 Breast Tum BC_γδ-T.3 |
| **IL17A** | | | | | | + |
| **IFNγ** | | | + | + | + | |
| **Antigen presentation** | | | + | | | |
| **Cytotoxicity** | | | + | ++ | + | |
| **Innate** | | | + | + | + | |
| **Annotation** | Naive or immature | Recent thymus emigrant | Absent in healthy blood/ associated with good outcome | Classical δ2 anti-bacterial/ anti-viral | | |

**Figure 3.16:** Marker genes and refined classification of γδ-T cell subtypes suggested from the data in this chapter. **(A)** Feature plots showing expression of suggested cluster defining markers in the γδ-T subtypes in PBMCs (top row) and BC (bottom row). Gray indicates low expression and purple indicates high expression. **(B)** Table summarizing the proposed refinement of subtype classification of γδ-T cells supported by the scRNA-seq data from this study.

## 3.4 Discussion

γδ-T cells are a very heterogeneous and poorly understood subset of T cells with untapped potential for exploitation for immunotherapy. Currently, subtypes of γδ-T are defined by their TCR sequences and in particular, the two well-defined ones from the circulation, namely *TRDV1* and *TRDV2* (Davey et al. (2018), Hayday (2019)). The data and analysis presented in this chapter have revealed that in the human peripheral blood, despite having the same TCR sequence, there are various gene expression-based cell states which are specialised for different functions.

Here I have used scRNA-seq to unbiasedly identify transcriptionally distinct subtypes and markers of γδ-T cells from human blood and breast tumour samples. This study is significant because 1) it identified new markers that will help sort γδ-T cell subtypes so they can be further studied; 2) it identified previously unappreciated heterogeneity in both the δ1 and δ2 subtypes; and 3) it identified a δ1-like subtype in breast cancer that is associated with better survival.

The results are consistent with the previously described *CD16* - *CD28* axis in the δ2 clusters (Ryan et al., 2016); however, the new data presented here suggests that *CD28* is also present in δ1 subtypes and therefore not a good marker. *GPR56* was identified and validated as an alternative marker to *CD16* and *CXCR6* as a more specific marker of the two non-*CD16* δ2 subtypes (**Figure 3.9**).

In PBMC, the two δ1 subtypes have subtle differences but one is immature while the other is recent thymus emigrant and has residual cytotoxic and antigen potential. Of the three PBMC δ2 subtypes, one is marked by *CCR6* and is an IL17A producer whilst the other two (*CD16*+ and *CCR7*-*CD16*-) are IFNγ producers. The *CCR6*+ and the *CD16*+ subtypes are detected in both PBMC and BC. The one of the two IFNγ producer clusters that is marked by *CD16*+

has higher antigen presentation and lower cytotoxic potential than the one that is *CCR6-CD16-*.

In BC, three subtypes were identified: an *IL17A* producer δ2 subtype (also marked by *CCR6*), an IFNγ producer and highly cytotoxic subtype marked by *CD16* (and thus likely a δ2 subtype) and a δ1 subtype which was also an IFNγ producer with high antigen presentation capacity (marked by *NCAM1*). Although previous work showed that γδ-T cell abundance is correlated with better outcome for several types of cancer (Gentles et al., 2015), here we see that, at least for the breast cancer cases considered, it is only the presence of one specific subtype that is associated with better outcome. Mapping of gene signatures suggests that this subtype has no equivalent in PBMC.

The subtype of γδ-T cell defined in this chapter that associates with better overall survival of breast cancer patients may be very similar or equivalent to the one recently found by Wu et al (Wu et al., 2019) who also identify a δ1, *IL17A* low/negative, *IFNγ+* subtype that associates with better overall survival of TNBC patients. While *TRDV1* is absent from the BC data, clear enrichment of *TRGV4* suggests that this subpopulation is δ1. Although Wu et al (Wu et al., 2019) performed a battery of functional assays on γδ-T cells in the in vitro setting, the results presented in this chapter can help better understand this subtype by defining differentially expressed genes that could be used as potential biomarkers for isolation or targeting of this cell subpopulation.

A limitation of this study is that it does not capture very rare γδ-T subtypes and the full extent of *TRD-TRG* pairing. Moreover, due to the cost of single cell RNA-sequencing and the difficulty of obtaining fresh tissue this study is based on a limited number of patients. Although it would be interesting to compare γδ-T cells between blood and breast cancer samples from the same individuals, it was not possible to obtain PBMC samples from the two breast cancer patients

(BC1 and BC2). Additionally, none of the two breast cancer samples included NK cells. It would be important to obtain datasets that consist of both It would be important to analyse datasets that include NK cells which were absent from both breast cancer samples, as these are expected to be the most transcriptionally and functionally similar cell type to δ-T and NK cells to describe the transcriptional and functional differences between these two similar immune cell types.

Future directions should involve validating these findings in additional samples, functionally characterizing the *NCAM1+* subtype, and determining if presence of the *NCAM1+* subtype also correlates with better survival in other cancers. Publicly available atlas-level data sets of breast cancer infiltrate immune cells, such as the data from Azizi et al. (2018) could be analysed to detect γδ-T cells, validate the presence of the subtypes defined here, look into the interactions of these subtypes with other cell types in the tumour microenvironment and evaluate the clinical relevance of the *NCAM1+* subtype. scID, presented in Chapter 2, can be a useful tool for querying these data sets.

The clinical relevance of the *NCAM1+* γδ-T subtype should be extended to a larger cohort. AutoGeneS (Hananeh and Theis, 2020) has been recently developed to use singe-cell extracted cell type signatures to perform deconvolution in bulk RNA-seq data and could be used to analyse more datasets from breast cancer studies in combination with the single-cell RNA-seq data presented in this chapter. Further endeavours to perform single cell RNA-sequencing of γδ-T cells from other tissues and tumour types are likely to be equally fruitful.

### 3.4.1 Conclusions

In summary, this chapter presents a large reference-level scRNA-seq dataset on blood γδ-T cells generated from three healthy donors and from two breast tumours. Analysis of this large dataset containing transcriptomes from a total of

7,000 γδ-T cells identified multiple novel subsets and marker genes for both δ1 and δ2 subpopulations, including a δ1 breast tumour infiltrating γδ-T cell subtype that is absent in PBMC and that is associated with favourable overall survival of breast cancer patients. This data suggests a refined classification of γδ-T cells shown in **Figure 3.16**.

Taken together, the data and results presented in this chapter have contributed to a better understanding of the functional diversity of γδ-T and will serve as a valuable resource for the community that can be mined to identify markers useful for isolating novel subsets to be able to further interrogate their functions.

# Chapter 4

# Demultiplexing donor identities in pooled single-cell RNA-seq data

## 4.1 Introduction

In previous chapters, I discussed the importance of comparing equivalent cell types across different conditions in order to understand their roles in various diseases and their response to treatment. We also saw that the technical variation introduced when samples are processed separately, also known as batch effect, is much stronger than the biological variation and leads to cells being grouped by these technical characteristics rather than biological similarity. Even when alignment or mapping is used to identify equivalent cell types across datasets, measured gene expression is still confounded by batch and does not allow further comparison of the cells from the different datasets, such as differential gene expression analysis or pseudotemporal ordering.

In order to be able to disentangle biological from technical variation, pooled experiments are performed, where cells from different samples are mixed and processed together in a single RNA-seq experiment. An additional advantage of clusters of pooled cells is the ability to show whether there are differences in proportions of cells between individuals. This could explain differences in disease progression and response to therapy observed between individuals with the same disease. Finally, pooling samples prior to single-cell RNA-sequencing can reduce the per-sample library cost, and thus allow for single-cell RNA-sequencing experiments at a population scale.

While most droplet-based methods become cost effective only when a very large number of cells is pooled and a very large number of cells will be required in population-scale scRNA-seq experiments, increased numbers of cells loaded lead to an increased proportion of multiplets obtained as discussed in Chapter 1. Additionally, in order to be able to compare equivalent cell types across samples, we need to be able to know the sample of origin of each cell after pooling for further analysis. Both experimental and computational solutions have been developed to tackle these problems.

### 4.1.1 Experimental demultiplexing of sample identity

Recent experimental protocols have been developed to enable tracking of sample identity for each cell prior to pooling cells from multiple samples. SPLiT-Seq (Rosenberg et al., 2018) achieves this with an extra barcoding step prior to cell and DNA tagging as follows. Dissociated cells from each sample are placed into a 96-well plate, one sample per well with a unique barcoded primer. This adds an extra tag to each transcript that allows tracking of its sample of origin. Such protocols, however, increase the complexity of the experiment that might add extra stress to the cells as well as increase the cost of the experiment. Additionally, the increased length of the barcode sequence, i.e a sample barcode, a cell barcode

and a UMI, leads to a decreased length of the captured transcripts that can be sequenced, resulting in even sparser single cell RNA-seq data.

Other methods, such as CITE-Seq (Stoeckius et al., 2018), assign sample labels to the cells using oligo-tagged antibodies that target sample-specific cell surface proteins. Besides the fact that obtaining proteomics data increases the cost of the experiment, the most important limitation of such approaches is the requirement for the existence of cell-surface proteins that are sample specific and highly expressed in all cells of a sample regardless of their cell type.

### 4.1.2 Computational demultiplexing of sample identity

An alternative approach to the above experimental methods is computational demultiplexing of sample identity using genetic information. Methods have been developed to cluster the cells based on donor-specific genetic variants. These variants can either be available as external information from genotypic profiling of the donors, or be inferred from the single-cell RNA-sequencing data.

#### Demuxlet

One of the first such approaches was Demuxlet (Kang et al., 2018) that aimed to combine genetic variation information from single cell RNA seq data and genotype information from each donor to assign cells to donors and also identify simultaneously doublets formed of cells from two different donors. Demuxlet is using prior genotype information from each donor to calculate the probabilities of a cell originating from each donor given the observed variants in the cell's RNA reads accounting for the dropout rate and the population allele frequency of the alternative allele. A mixture model is also used to calculate the probability of a cell being actually a doublet originating from two individuals, for each pairwise

combination of individuals. All likelihoods are finally compared to annotate the cell.

### Cardelino

Similarly, Cardelino (McCarthy et al., 2018) uses bulk or single-cell DNA-sequencing data from the different donors to infer a clonal tree from frequency and co-occurrence of genetic variants and then assign pooled cells from single cell RNA-seq data to the corresponding donors using a Bayesian mixture model.

Both Cardelino and Demuxlte are highly accurate, especially in presence of a high number of pooled donors or when donors are genetically close. However, they are restricted to samples that originate from previously genotyped individuals and thus, are not applicable to already publicly available datasets. Additionally, the cost of genotyping the samples of all pooled donors has to be added on top of the single-cell sequencing cost of a study.

### Vireo

Vireo (Huang et al., 2019) on the other hand, is able to demultiplex scRNA-seq data without the need for prior genotype information. Given the observed reference and alternative counts of known genetic variants (e.g. 1000 Genomes Project) in each single cell, Vireo implements a variational Bayesian inference model to estimate both the donor identity of each cell and the genotype of each donor simultaneously.

In both Vireo and Cardelino, genomic variants are extracted from single-cell RNA-sequencing data using `cellSNP` (Huang, 2019).

**scSplit**

Another recently published method that does not require genotype information or a list of common variants is scSplit (Xu et al., 2019). scSplit uses the counts of reference and alternative alleles for each genomic position and each cell and initializes the clusters' centroids with k-means clustering. Each cluster's centroid is then represented as the alternative allele frequency. The cells are assigned to clusters/donors based on their probability of being observed from each sample and the centroids are updated iteratively until convergence, using the Expectation-Maximisation algorithm.

### 4.1.3 Computational challenges of demultiplexing

In theory, a single variant is sufficient to distinguish between cells from two donors, but this requires that this variant is captured in all cells of the donor. However, Single Nucleotide Polymorphism (SNP) identification in single-cell RNA-seq data poses several technical challenges. The main limitation is the low number of SNPs we can detect for each cell. This is due to the fact that we can only detect exonic SNPs and only in the expressed genes of a cell which is a small proportion of the total number genes. Additionally, most methods only capture the 3'UTR ends of the transcriptome, further reducing the probability of observing a SNP due to the low number of nucleotides sequenced. Finally, on top of that is added the high sparsity of the data due to dropouts, i.e. genes that are not captured although they are expressed in the cell.

All these effects result in a much smaller number of SNPs being captured in each cell compared to the actual SNPs we would observe from that donor from bulk or even single-cell DNA-sequencing data. Despite the additional sparsity of the SNP data compared to the gene expression data, the dimensionality of the data remains very high. Dimension reduction is required due to the lack of sufficient

number of samples even when thousands of cells can be sequenced. Finally, as technology improves and cost decreases, we expect more and more cells to be able to be processed simultaneously and more and more individuals to be multiplexed, which requires a fast and scalable solution.

### 4.1.4 Aim and structure of this chapter

In this chapter I will be focusing on the application of Deep Learning methods to identify the donor identity of cells in a pooled single-cell RNA-sequencing dataset without prior genotype information from the donors. More specifically, I will be showing a pipeline for extracting SNPs from the FASTQ files of the single-cell RNA-seq data and clustering this data to group cells by donor identity. I will be using non-negative matrix factorisation, in order to obtain a lower dimensional representation of the original data that can additionally show the specific variants used to classify the cells into donors. This information is very useful for validating the results.

## 4.2 Methods

### 4.2.1 Deep Learning

Deep learning is a class of machine learning methods inspired from the structure and function of the brain, that aims to solve a problem by extracting higher level features from raw input data using multiple layers of non-linear transformations (Lecun et al., 2015). Similar to the structure of the nervous system with neurons being connected to each other and passing information, deep learning models consist of interconnected nodes that exchange information. These nodes are however organised in layers with each layer accepting information only from the previous layer and passing it on to the next layer. The first layer is the input and the last layer is the output of the model and these are the only observable layers. Any intermediate layer is called "hidden" (**Figure 4.1**)

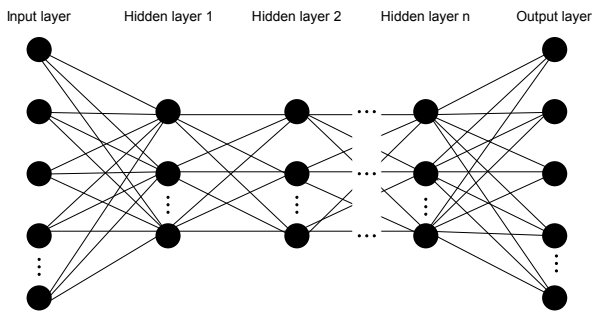



**Figure 4.1:** General structure of a deep learning network. Each layer may have a different number of nodes.

Deep learning has the ability to process natural data in their raw format. Unlike conventional machine-learning methods that require a well-designed feature extraction strategy, deep learning methods can automatically extract informative features from the data and learn very complex functions to transform them to higher representations by combining multiple non-linear transformations. Such higher representations have the ability to enhance the effect of variables that

are important for classification and suppress irrelevant variations. In high-dimensional data, deep learning is able to discover complex structures. Another advantage of Deep Learning compared to conventional machine learning methods is the ability to improve their performance when the amount of data available increases. Other machine learning methods, on the other hand, reach a plateau of the performance after a specific amount of input data. This is a very important advantage in the area of single cell data as new experimental protocols enable the generation of a vast amount of publicly available single-cell RNA-sequencing datasets.

Deep learning methods can be either supervised or unsupervised. Supervised learning builds classifiers by training a deep network model with a vast amount of labelled datasets and then the trained model is used to classify new unlabelled datasets. However, supervised algorithms have several drawbacks. First, they require a vast amount of labelled data for training that is not always possible to obtain. Additionally, there is a risk of the model being bound by the biases in the training data (overfitting). Unsupervised learning algorithms on the other hand can extract informative features from the testing data itself that can be used to group the data in an informative way. The idea of unsupervised learning is similar to human and animal learning that involves observation of the objects to understand their characteristics.

Deep learning methods have led to application breakthroughs in a wide range of science and business fields, from image and natural language processing to analysing particle accelerator data. In the field of genomics and medicine, deep learning methods have been used to predict the effect of non-coding variants (Zhou and Troyanskaya, 2015), the effect of mutations in gene expression and cancer (Luo et al., 2019) or the effect of drugs (Kalinin et al., 2018).

### 4.2.2 Non-negative matrix factorisation

Non-negative Matrix Factorisation (NMF) is widely used in applications aiming to separate non-negative sources from complex mixtures. The basic model of NMF is a factorisation of a non-negative input matrix $\mathbf{X}$ into two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$, each of them being of lower rank than $\mathbf{X}$, such that

$$\mathbf{X} = \mathbf{WH} \tag{4.1}$$

Non-negative refers to all elements of the factor matrices $\mathbf{W}$ and $\mathbf{H}$ being equal to or greater than zero.

The optimal functions are the solutions to the constrained optimisation problem

$$(\hat{\mathbf{W}}, \hat{\mathbf{H}}) = \underset{\mathbf{W},\mathbf{H}\geq 0}{\arg\min}\, f(\mathbf{W}, \mathbf{H}) = \|\mathbf{X} - \mathbf{WH}\|_F^2 \tag{4.2}$$

where $\|\cdot\|_F^2$ indicates the Frobenius norm and $f$ is the squared Euclidean distance function between the observed data $\mathbf{X}$ and the estimated $\mathbf{WH}$.

NMF is conceptually similar to dimension reduction methods, such as PCA and Independent Component Analysis (ICA), however, NMF produces a sparse and parts representation of the high-dimensional data that leads to easier interpretation of the factor matrices $\mathbf{W}$ and $\mathbf{H}$ as physical building blocks of the input $\mathbf{X}$ (Montesdeoca et al., 2019).

Similar to PCA and ICA, NMF is a deterministic mapping between the input and the latent space. However, for sparse data such as single-cell RNA-sequencing data, a probabilistic approach is more suitable in order to represent the data by approximating the underlying probability distribution.

### 4.2.3 Probabilistic Non-negative Matrix Factorisation with variational autoencoder

Several probabilistic extensions to non-negative matrix factorisation have been proposed (Bayar et al. (2014), Heinz (2014), Mohammadiha et al. (2013), Cemgil (2009), Schmidt et al. (2009)) that show advantages over deterministic NMF in applications with noisy data, such as clustering of DNA microarray data (Bayar et al., 2014). However, all above methods require the selection of appropriate priors which require computationally intensive techniques such as Monte Carlo to compute the full posterior.

Montesdeoca et al. (2019) on the other hand, suggest the use of Variational Autoencoder (VAE) for PNMF (PAE_NMF). VAEs are probabilistic models that utilise the autoencoder (AE) framework for a neural network to find probabilistic mappings from the input to the hidden (latent) layers and from the hidden layers to the output. AEs consist of an encoder part that compresses the input into a fewer dimensions layer and a decoder part that tries to reconstruct the input from the compressed data. This forces the network to retain only informative features of the data which works as denoising. AEs aim to only approximate the input instead of perfectly reconstructing it, enabling the network to learn useful properties of the data by ignoring signal noise, without the risk of overfitting. In a simple autoencoder network shown in **Figure 4.2**, the coefficients matrix $\mathbf{H}$ is the latent representation of the input $\mathbf{X}$ and the final weights should be non-negative with an identity activation function such that the output $\hat{\mathbf{X}} = \mathbf{WH}$ and $\hat{\mathbf{X}} \approx \mathbf{X}$.

The combination of the VAE framework and NMF sets a non-negative constraint on the latent space of the autoencoder leading to interpretable results while maintaining a higher degree of flexibility on the posterior distribution and automatic regularisation. Here we have implemented the model suggested by

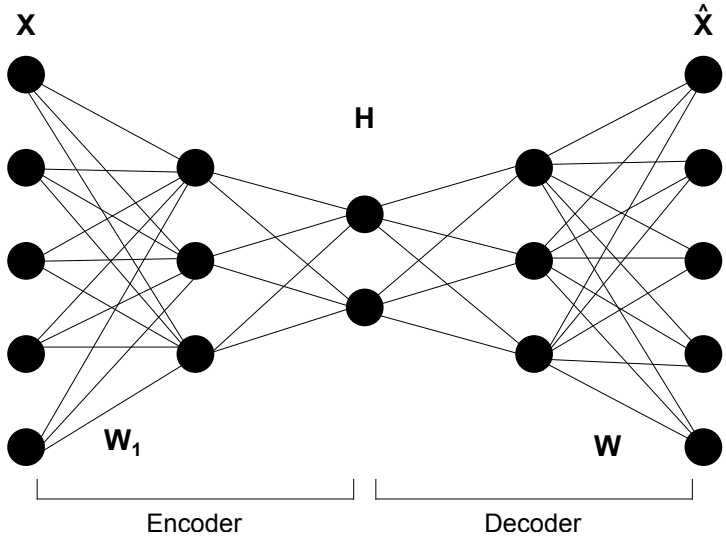



**Figure 4.2:** Simple autoencoder network for non-negative matrix factorisation. The encoder part compresses the input into a fewer dimensions layer (here shown in two dimensions) and the decoder part tries to reconstruct the input from the compressed data.

Montesdeoca et al. (2019) to identify the sample identity of cells based on observed single-nucleotide polymorphisms (SNPs) without genotypic information of the individuals. More specifically, given a non-negative matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, with $m$ being the number of observed SNPs and $n$ being the total number of cells in the dataset, we seek to deconvolve this to a factor matrix $\mathbf{W} \in \mathbb{R}^{m \times l}$ representing the load of each SNP in each sample identity and a coefficients matrix $\mathbf{H} \in \mathbb{R}^{l \times n}$ representing the sample identity of each cell, with $l$ being the number of different individuals in the pooled scRNA-seq data by minimising:

$$\frac{1}{2}\left\|\mathbf{X} - \hat{\mathbf{X}}\right\|_F^2 + D_{KL}(q_\phi(H|X)\|p(H)) \tag{4.3}$$

The first term of equation 4.3 represents the reconstruction error between the input and the estimated output measured by the Frobenius norm and the second term represents the KL-divergence between the prior ($p(H)$) and the posterior

($q_\phi(H|X)$) and acts as regularisation term forcing the posterior to remain close to the prior.

As shown in **Figure 4.3**, the encoder is used to estimate the parameters $\mathbf{k}$ and $\boldsymbol{\lambda}$ of the posterior distribution and the coefficients matrix $\mathbf{H}$ is obtained by sampling from the posterior distribution. Sampling however from the posterior distribution does not allow calculation of the derivatives $\frac{\partial H_i}{\partial k_i}$ and $\frac{\partial H_i}{\partial \lambda_i}$ for each dimension $i$ required in the backpropagation step. Thus variational autoencoders use an extra input node $\boldsymbol{\epsilon}$ to introduce the stochasticity in the middle of the network without needing to differentiate it through, making the $\mathbf{H}$ node deterministic. Finally, the decoder part of the network is used to reconstruct the input from $\mathbf{H}$ such that the estimated output $\hat{\mathbf{X}} \approx \mathbf{X}$

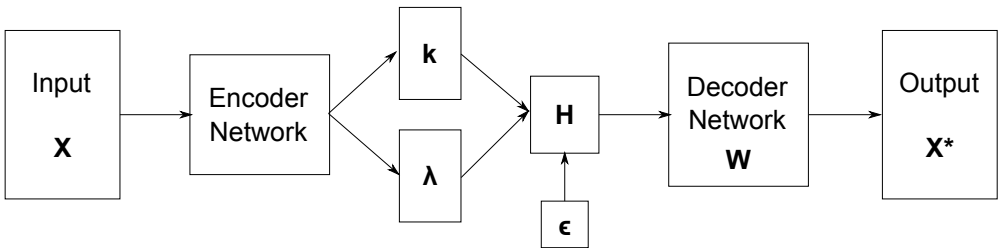



**Figure 4.3:** Schematic representation of the architecture of a variational autoencoder for non-negative matrix factorisation. Adapted from Montesdeoca et al. (2019).

Standard VAEs use Gaussian distributions. However, with the restriction of non-negativity we require a distribution that falls towards zero when $x$ becomes large and has a Probability Density Function (PDF) of zero below $x = 0$. Although other distributions, such as NB, ZINB and gamma distribution have been used for modelling genomic data (Eraslan et al., 2019) and for NMF (Squires et al., 2017), for the reparametrisation trick and inverse transform sampling a Weibull distribution is simpler. The PDF of a Weibull distribution is defined as:

$$f(x) = \begin{cases} \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} exp(-(\frac{x}{\lambda})^k), \text{if } x \geq 0 \\ 0, \text{if } x < 0 \end{cases}$$

Thus, for sampling from the Weibull distribution we need the inverse cumulative function:

$$\mathbf{H} = C^{-}1(\boldsymbol{\epsilon}) = \boldsymbol{\lambda}(-ln(\boldsymbol{\epsilon}))^{1/k} \tag{4.4}$$

where $\epsilon$ is a random variable sampled from a uniform distribution $\epsilon \sim \mathcal{N}(0, 1)$.

The KL-divergence between two Weibull distributions for the second term of the optimisation problem of Equation 4.3 is given by:

$$D_{KL} = log(\frac{k_1}{\lambda_1^{k_1}} - log(\frac{k_2}{\lambda_2^{k_2}} + (k_1 - k_2)[log(\lambda_1) - \frac{\gamma}{k_1}] + (\frac{\lambda_1}{\lambda_2})^{k_2}\Gamma(\frac{k_2}{k_1} + 1) - 1 \tag{4.5}$$

where $k_1$ and $\lambda_1$ are the parameters of the posterior distribution, $k_2$ and $\lambda_2$ are the parameters of the prior distribution, $\Gamma$ is the gamma function and $\gamma \approx 0.5572$ is the Euler-Mascheroni constant.

Having all the components of the network we can summarise the structure for a one-layer network as follows:

$$\boldsymbol{\lambda} = f(\mathbf{W}_{\boldsymbol{\lambda}}\mathbf{X}), \mathbf{k} = g(\mathbf{W}_{\mathbf{k}}X), \mathbf{H} = C^{-1}_{\mathbf{k},\boldsymbol{\lambda}}(\boldsymbol{\epsilon}), \hat{\mathbf{X}} = \mathbf{W}_f\mathbf{H} \tag{4.6}$$

where $f$ and $g$ are element-wise non-linear functions defined by the neural network and the inverse cumulative function $C^{-1}_{\mathbf{k},\boldsymbol{\lambda}}$ also works element-wise.

The PAE_NMF model was implemented in python using the open source machine learning library PyTorch (PyTorch, 2016).

**Extraction of donor-specific SNPs from W**

As previously mentioned, **W** is an $m \times l$ matrix representing the load of $m$ SNPs in the identity of each of the $l$ donors. For example, for a dataset with two donors (i.e. $l = 2$), if $w_{i1} \gg w_{i2}$ for a SNP $i \in 1, 2, ..., m$ then SNP $i$ is present in sample 1 and absent in sample 2. For these two-donor cases, I calculate the difference $d_i = w_{i1} - w_{i2}$ for all SNPs. SNPs with positive $d_i$ are defining donor 1 and SNPs with negative $d_i$ are defining donor 2. To obtain SNPs with high contribution to each donor's identity, these SNPs are sorted by $d$ and top positive and negative SNPs are selected.

As expected, the majority of the SNPs will be noisy and will not be strongly associated to any donor. To identify these SNPs we can fit a linear regression in the relationship between $w_{i1}$ and $w_{i2}$ and filter out SNPs that are within one standard deviation from the fitted line.

### 4.2.4 Extraction of single nucleotide polymorphisms from single-cell RNA-seq data

For identifying single nucleotide variants in the single cell RNA sequencing data I designed the following pipeline using a BAM file as input. The pipeline uses the BAM fields and tags shown in **Table 4.1**. Detailed information on these fields and tags can be found in the SAM format specification document (Li et al., 2009).

The MD tag is complementary to the CIGAR string as it can distinguish between matches and mismatches that are both denoted as "M" in the CIGAR string, thus why it used here. For example, MD = "10A25^4" indicates that the first 10 bases are identical between the reference and the query sequence, then the reference has an A while the query sequence has whatever is reported in the `SEQ` field, the next 25 bases are identical and then there is a deletion in the query sequence indicated

**Table 4.1:** Fields and tags obtained from BAM file for extraction of single nucleotide polymorphisms from single-cell RNA-seq data.

| Field/Tag | Description |
|---|---|
| `RNAME` | Reference chromosome. |
| `POS` | Leftmost mapping position of the reference genome. |
| `MAPQ` | Mapping quality. Maximum is 255 for alignment with STAR |
| `FLAG` | Bitwise describing several features of the aligned sequence. Used to infer whether the sequence was aligned to the forward or reverse strand |
| `CB/CR` | Cell barcode raw (`CB`) or error-corrected and confirmed against a list of true barcodes (`CR`). `CR` is preferred if available. |
| `UB/UR` | Unique Molecular Identifier (UMI) raw (`UB`) or error-corrected (`UR`). `UR` is preferred if available. |
| `NM` | Number of mismatches. Require exactly 1 mismatch to retain read. |
| `MD` | String for mismatching positions complementary to the `CIGAR` string. |

by the "^" symbol. Finally the last 4 bases are identical to the reference. To obtain our variants matrix we only retain reads with a single mismatch, thus we keep only reads with MD consisting of a single letter and numbers.

- **Step 1**: Only retain reads with mapping quality equal to 255 for STAR alignment to ensure that observed variants are not due to sequencing errors.
- **Step 2**: Only retain reads in the forward strand as defined by transcriptional orientation.
- **Step 3**: Retain reads of filtered cells based on their associated barcode for Chromium 10X data.

- **Step 4**: Only retain reads with exactly one mutation (tag `NM` = 1).
- **Step 5**: Of these reads remove those that have deletions (tag `MD` contains "^" symbol).
- **Step 6**: Filter out mutations that are outside of a gene coding region as these are likely to be alignment errors.

A binary matrix showing whether a specific SNP (rows) is present (value of 1) or absent (value of 0) in each cell (columns) will be used as input to the PAE_NMF model.

Preliminary comparison of the above binary matrix and the identified variants using cellSNP as input to PAE_NMF did not show significant differences in the classification accuracy. This is due to PAE_NMF not using allele frequencies that are additionally calculated with cellSNP but not with the above pipeline. Since the above method is marginally faster to run compared to cellSNP, it was selected for all tests shown in the next section.

## 4.3 Results

It is known that performance of variational autoencoders is highly dependent on the choice of hyperparameters (Hu and Greene, 2019). The main hyperparameters of this method are the number of layers, the number of hidden units (nodes) in each layer, the number of epochs, the batch size and the learning rate. The number of layers and hidden nodes depends on the size of the data. Too many layers and nodes could lead to overfitting, while too few could lead to high bias. For our high dimensional data, we do not have sufficient samples to estimate internal weights of very deep wide networks. I have chosen to use two layers and test the accuracy of the classification for different numbers of hidden nodes. Thus the architecture of the network will be M - $d_1$ - $d_2$ - k - $d_2$ - $d_1$ - M, where M is the dimension of the input data, i.e. the number of SNPs, $d_1$ and $d_2$ and the number of nodes in layers 1 and 2, respectively, and k is the size of the bottleneck of the network that should be equal to the total number of donors in the data.

The number of epochs is the number of times the training data is shown to the model. High number of epochs might lead to overfitting and low number might lead to poor training of the model. Different values will be tested to study the effect of the number of epochs on the classification accuracy for each dataset. In every epoch, the model is trained with different batches (sets) of samples (cells). The batch size indicates the number of samples used in each batch. Small batches can lead to unstable learning as feature patterns will be less repeating between batches. Large batches, on the other hand, can lead to very slow learning as there will be small variation between batches. Finally, the learning rate is automatically adjusted during each iteration using the Adam optimizer (Kingma and Ba, 2015) in order to avoid getting stuck in local optima or not converging to the minima.

Besides the hyperparameters of the variational autoencoder, I will be testing the

effect of SNP filtering on the classification accuracy. SNPs will be filtered based on their prevalence in the data. SNPs that are present in very few cells are not expected to be useful for the classification of the cells and removing them can help slightly reduce the dimensionality of the data.

Finally, I will be comparing the classification accuracy of the PAE_NMF model to that of the previously mentioned methods.

### 4.3.1 Test Case: 50%:50% Jurkat:293T cell mixture

A simple example to test the PAE_NMF method and compare it to existing methods is a mix of 293T and Jurkat cells provided from 10X Genomics. These are two distinct cell lines originating from different tissues with little or no intra-sample heterogeneity of gene expression. The 293T cell line originates from human embryonic kidney cells (Graham et al., 1977), while the Jurkat cell line consists of T cells from peripheral blood of a 14-year old boy with acute lymphoblastic leukaemia (Schneider et al., 1977). Thus, it should be easy to detect genetic variants that discriminate the two samples. "Ground truth" labels have been obtained from the output of demuxlet (demuxlet, 2017). According to these labels, of the 500 cells, 240 (48%) cells are 293T cells, 218 (43.6%) cells are Jurkat cells and 42 (8.4%) cells are doublets.

Using the previously described method for extraction of variants from the `BAM` file, a total of 11,317 variants were identified. Of these variants, 4288 (37.9%) are only observed in a single cell and a cell has on average 378 SNPs. **Figure 4.4 A** shows the distribution of number of SNPs per cell and **Figure 4.4 B** shows the distribution of the SNPs' prevalence.

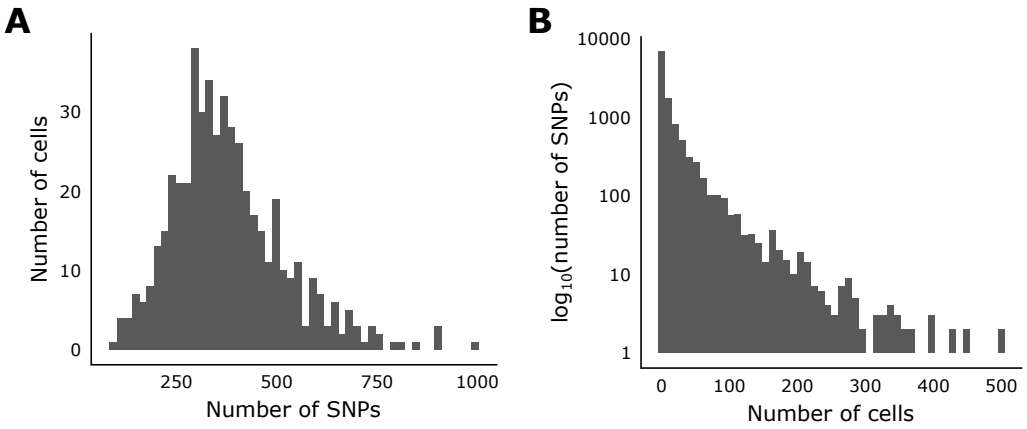



**Figure 4.4:** Distribution of identified SNPs. **(A)** Histogram showing number of identified SNPs per cell. **(B)** Histogram showing frequency of each identified SNP. $x$-axis shows number of cells a SNP is present and $y$-axis shows the number of SNPs in $log_{10}$ scale. SNP, single nucleotide polymorphism

### Effect of SNP filtering in classification accuracy

To reduce the dimensions of the variants matrix and remove noisy SNPs, I filtered SNPs that are present in very few cells. Based on the distribution of SNPs' frequency (**Figure 4.4 B**), there is no obvious threshold to select (e.g. the distribution being bimodal). Thus, I tried different thresholds of minimum number of cells per SNP from the set of $\{1, 2, 3, 4, 5, 10, 15, 20, 25\}$. **Figure 4.5** shows the effect of SNP filtering in the classification accuracy for different combinations of hyperparameters. The classification accuracy as measured by the ARI (see Section 2.2.6) between the "ground truth" and the PAE_NFM predicted labels is shown on the $y$-axis. Each boxplot shows the ARI for a specific threshold of SNPs' prevalence (shown on the $x$-axis) and for different values of the previously mentioned hyperparameters. The distribution of ARI within each group is very wide, ranging from very poor to very accurate classification. However, there is no significant difference in the accuracy between the different SNP filtering thresholds.

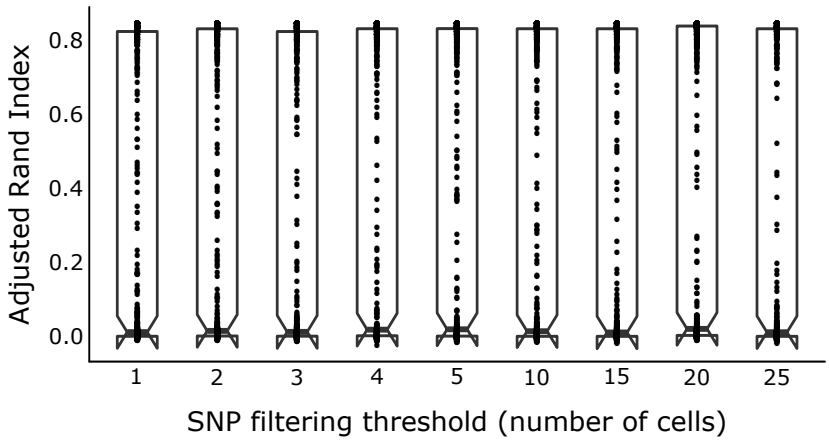



**Figure 4.5:** Effect of the SNP filtering threshold on the classification accuracy of PAE_NMF. Boxplot showing the ARI ($y$-axis) for different filtering thresholds of SNPs based on their prevalence ($x$-axis). Each dot represents the ARI for the threshold indicated on the $x$-axis and a different set of hyperparameters. ARI, Adjusted Rand Index

### Effect of depth of network

I next sought to test the effect of the number of nodes of each hidden layer in the classification accuracy. Using too few nodes is expected to result in the model not capturing well the true sources of variance in the data, known as underfitting. On the other hand, too many nodes can result in overfitting and inability of the model to remove noise from the data. However, there is no theoretically proven rule for selecting an appropriate number of nodes. Thus, I experimented with different numbers of nodes ranging between 10 and 150 while also changing the other hyperparameters of the model.

**Figure 4.6** shows the ARI of the classification of these models for different numbers of nodes in layer 1 (grey) and layer 2 (yellow). Each dot represents the average ARI of multiple models with the same number of nodes (indicated on the $x$-axis) and different values of the other hyperparameters and the error bars indicate the 95% confidence interval. Although there is a trend of increasing

ARI with decreasing number of nodes in layer 1 and with increasing number of nodes in layer 2, there is no significant difference between the different options according to pairwise Wilcoxon signed rank tests.

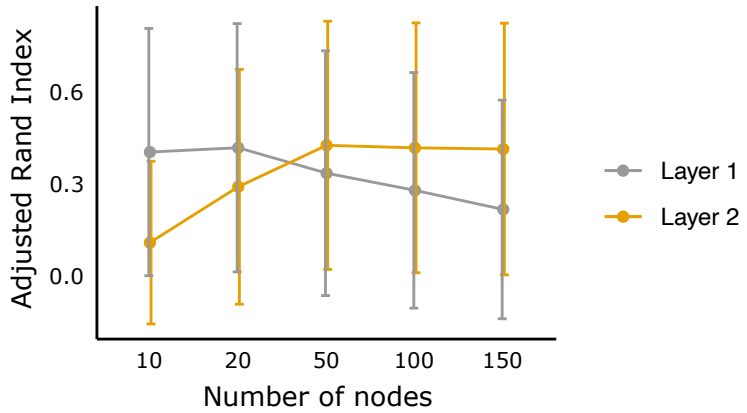



**Figure 4.6:** Effect of the number of nodes in each hidden layer on the classification accuracy of PAE_NMF for the Jurkat dataset. Scatter plot showing the ARI ($y$-axis) for different number of nodes ($x$-axis) of layer 1 (grey) and layer 2 (yellow). Dots show average ARI of multiple models with the same number of nodes and different values of the other hyperparameters (number of epochs and batch size) and error bars indicate 95% confidence interval. ARI, Adjusted Rand Index.

### Effect of number of epochs in the classification accuracy

As expected using an increased number of epochs leads to increased classification accuracy. **Figure 4.7** shows the classification accuracy of the network with different hyperparameters. A low number of epochs often resulted in poor classifications. The ARI on average is constantly improving with an increased number of epochs. However, there are still models with very poor performance after many iterations. Thus, there is no guarantee that any model will result in an accurate classification after 50, 75 or 100 epochs.

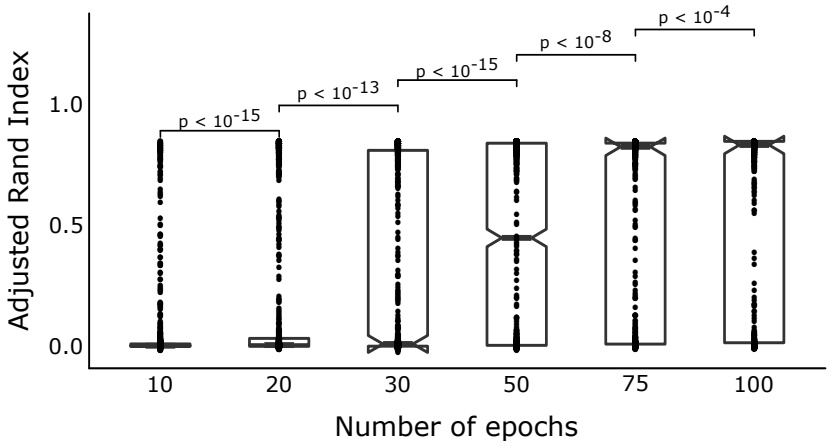



**Figure 4.7:** Effect of the number of epochs on the classification accuracy of PAE_NMF for the Jurkat dataset. Boxplot showing the ARI ($y$-axis) for different numbers of epochs ($x$-axis). Each dot represents the ARI for the number of epochs indicated on the $x$-axis and a different set of hyperparameters. ARI, Adjusted Rand Index

### Effect of batch size in the classification accuracy

Methods that use stochastic gradient descent operate in a small-batch regime, since use of larger batches has shown poor ability to generalize (Keskar et al., 2019). Small batch sizes, on the other hand offer a regularizing effect (Wilson and Martinez, 2003). This could be due to noise induction that in non-convex optimisation problems helps the model escape saddle points during training and bad local minima (Ge et al., 2015). Here I tested the accuracy of classification for batch sizes ranging between 4 and 128 (**Figure 4.8**). As expected, increased batch size leads to more models failing to accurately classify the cells, since there are no changes in the data presented to the model in each epoch that can help the model converge.

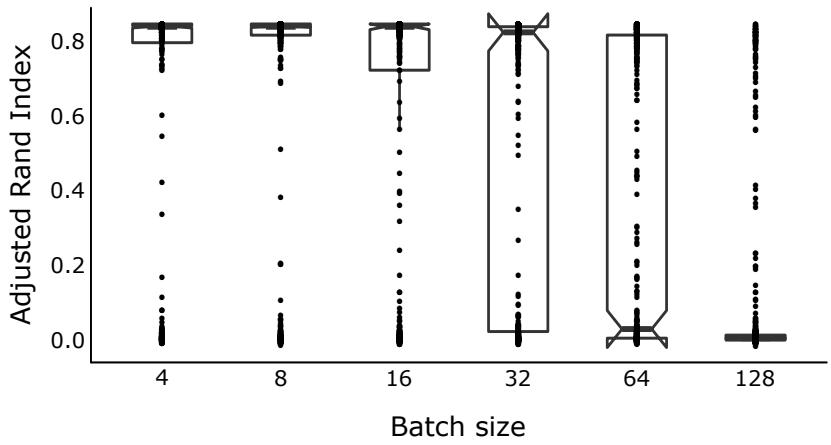



**Figure 4.8:** Effect of the batch size on the classification accuracy of PAE_NMF. Boxplot showing loss calculated after the last epoch for different batch sizes. Each dot represents the loss ($y$-axis) of a model with batch size indicated on the $x$-axis and a different set of hyperparameters (number of epochs, numbers of nodes, SNP filtering threshold). SNP, single nucleotide polymorphism

### Loss correlates with classification accuracy

The loss represents the difference between the observed variants matrix $\mathbf{X}$ and the estimated $\hat{\mathbf{X}} = \mathbf{WH}$. It is expected that the lower the loss the better is the classification accuracy of the model. This is confirmed by the classification accuracy of different models with batch size equal to 16 shown in **Figure 4.9**.

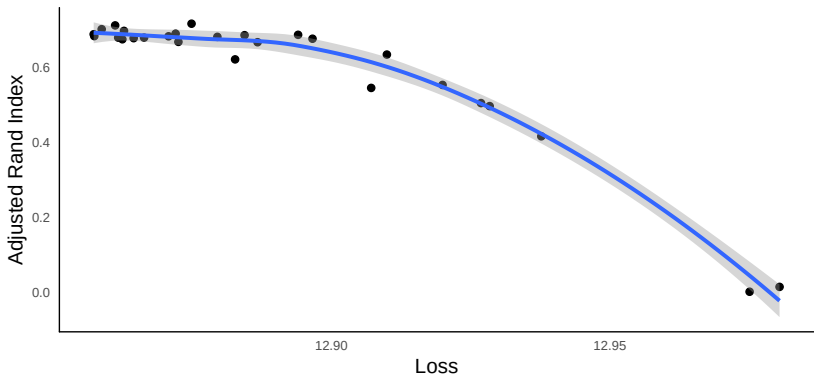



**Figure 4.9:** Relationship between loss and classification accuracy for batch size of 16.

However, the loss is highly dependent on the batch size (**Figure 4.10**). Although higher batch size leads to lower loss, it does not lead to better classification accuracy (**Figure 4.8**). Thus we can only compare the loss across models with the same batch size in order to decide which one is the most accurate.

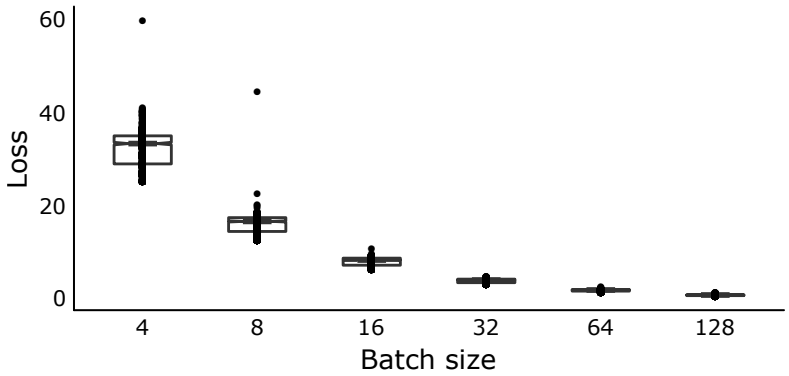



**Figure 4.10:** Boxplot showing loss calculated after the last epoch for different batch sizes. Each dot represents the loss ($y$-axis) of a model with batch size indicated on the $x$-axis and a different set of hyperparameters (number of epochs, numbers of nodes, SNP filtering threshold).

### Extraction of donor-specific variants for validation

Besides using the **H** matrix for obtaining the sample identities of the cells, non-negative matrix factorisation provides the ability to explore the **W** to find the genetic variants (SNPs) that are most important to the definition of each sample. These variants can be used for visualisation of the results as well as to guide further experimental validation.

**Figure 4.11 A** shows the top 10 variants per sample sorted by weight. The length of the bars represents the weight with positive values indicating enrichment in sample 1 and negative values indicating enrichment in sample 2. Variants are encoded as "chromosome_position". **Figure 4.11 B** heatmap shows the presence of these selected variants in all cells, grouped by sample identity according to the PAE_NMF method. Rows show the top 10 variants for sample 1 and sample 2

and columns show cells of sample 1 and sample 2. Red represents presence and blue represents absence of a variant in a cell. Given the dissimilarity of the two cell lines, all extracted variants are uniquely present in one of the samples with very low noise.

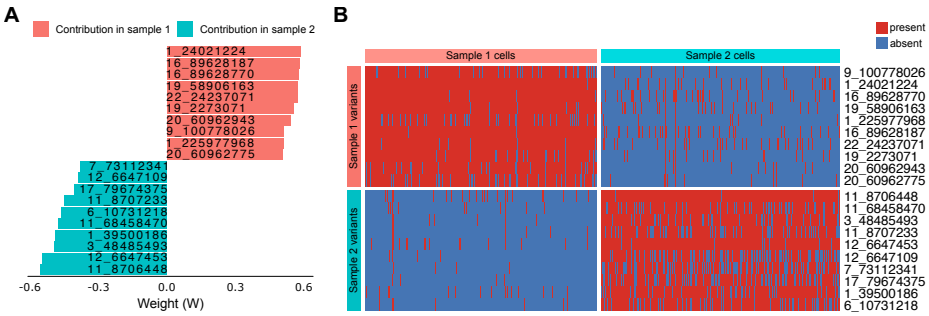



**Figure 4.11:** Extraction of sample-specific variants. **(A)** Barplot showing top 10 variants per donor sorted by weight. The length of the bars represents the weight with positive values indicating enrichment in donor 1 and negative values indicating enrichment in donor 2. Variants are encoded as "chromosome_position". **(B)** Heatmap showing the presence (red) or absence (blue) of the top 10 variants per donor (rows) in cells (columns) grouped by hierarchical clustering using Euclidean distance. Three main clusters are identified; cells corresponding to donor 1, cells corresponding to donor 2 and doublets enriched in SNPs from both donors.

### Doublet detection

Methods such as Demuxlet, Cardelino and Vireo use mixture models to calculate the probability of each cell originating from two individuals based on the presence of variants from different genotypes. scSplit includes an additional cluster to capture the doublets, based on the hypothesis that all doublets, regardless of the samples they originate from, will fall in the same cluster, separate from the singlets. One disadvantage of the method presented in this chapter is that it is not able to identify doublets as they are not an independent group of cells with unique variants.

To overcome this issue, we can use the extracted variants to detect cells that express variants from more than one sample. A simple solution is to use hierarchical clustering with Euclidean distance to group the cells into three clusters; one for each sample and one for the doublets. This results in a refined classification shown in **Figure 4.12**, where 243 cells are assigned to sample 1, 252 cells are assigned to sample 2 and 5 cells are identified as doublets. However, doublets consisting of two cells from the same sample cannot be identified with this approach or with any of the published methods.

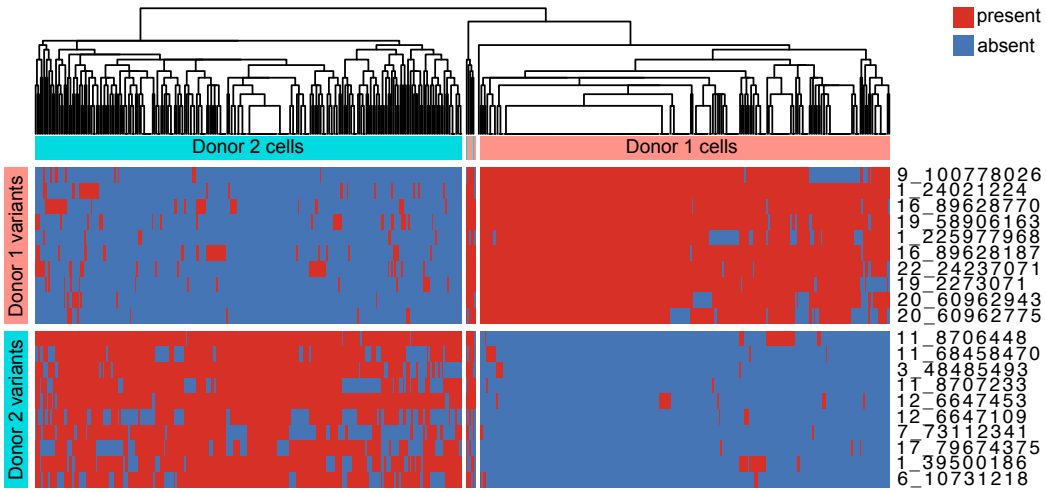



**Figure 4.12:** Doublet detection with hierarchical clustering. Heatmap showing the presence (red) or absence (blue) of the top 10 variants per sample (rows) in cells (columns) grouped by hierarchical clustering using Euclidean distance. Three main clusters are identified; cells corresponding to sample 1, cells corresponding to sample 2 and doublets enriched in SNPs from both samples.

### Comparison with existing methods

Next, I compared the accuracy of the classification of PAE_NMF with and without identification of doublets with that of vireo and scSplit. Demuxlet labels were used as "ground truth". **Figure 4.13 A** shows the ARI between the predicted labels from each method indicated on the $x$-axis and the labels obtained from demuxlet.

For PAE_NMF, six different results have been used, the ones with the lowest loss between all tests with equal batch size. The bar shows the average ARI from these tests and the error bar shows the 95% confidence interval.

The result of PAE_NMF with and without refinement is very similar to that of demuxlet and vireo. scSplit seems to have the best performance, however, almost 40% of the cells have been classified as doublets (**Figure 4.13 B**), which is much higher than the expectation and demuxlet's and vireo's prediction. The cells that have been identified as doublets with hierarchical clustering highly overlap with those selected as doublets by vireo and are a subset of those from demuxlet (**Figure 4.13 C**). The 27 doublet cells unique in demuxlet could be doublets from the same sample that are also considered by this method.

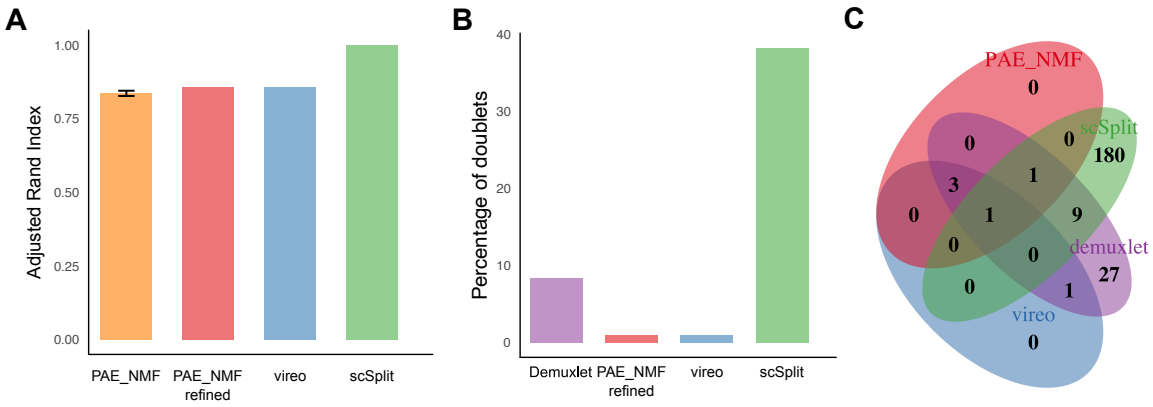



**Figure 4.13:** Comparison of PAE_NMF to vireo and demuxlet, using demuxlet labels as "ground truth". **(A)** Barplot showing Adjusted Rand Index ($y$-axis) between the predicted labels from each method indicated on the $x$-axis and demuxlet labels. Doublets have been removed from this comparison. For PAE_NMF, six different results have been used, the ones with the lowest loss between all tests with equal batch size. Bar shows the average ARI and the error bar shows the 95% confidence interval. "PAE_NMF_refined" is the result from the combination of PAE_NMF and hierarchical clustering using the top 10 variants per sample. **(B)** Barplot showing the percentage of identified doublets from each method. **(C)** Venn diagram showing overlap of the identified doublets between the four methods.

### 4.3.2 Application: Mix of γδ-T cells from donor 1 and CD3+ T-cells from donor 2

To confirm the above conclusions regarding the tuning of hyperparameters of the PAE_NMF model as well as its accuracy compared to other methods, I used a more complex dataset that consists of a mix of γδ-T cells from one donor and CD3+ T-cells from another donor. Although the number of donors is still low, there is a higher number of cells and there is additional gene expression heterogeneity within each donor's cells due to different cell types.

Peripheral blood γδ-T cells from one donor and peripheral blood CD3+ T-cells from a second donor were pooled before performing scRNA-seq. The obtained dataset consists of 6690 cells with an average of 863 expressed genes per cell. Unsupervised clustering of the gene expression data identified 10 clusters. Due to the different types of cells selected from each donor, we expect each cluster to consist of cells from a single donor, thus we are able to manually annotate the donor identities of the cells based on expression of known CD3-T and γδ-T cell markers. These labels will be used as "ground truth" to evaluate the classification accuracy of each demultiplexing method. Although γδ-T cells are a subpopulation of CD3+ T-cells, we do not expect to see a significant number of γδ-T cells from donor 2 due to their very low prevalence.

Based on expression of known T cell markers, i.e. *CD3D*, *CD4*, *CD8A* and *FOXP3*, and genes coding for the gamma and delta chains of the γδ-T cells, i.e. *TRDC*, *TRDV1*, *TRDV2*, *TRGV4* and *TRGV9*, that we saw in Chapter 3, we can obtain a manual classification. From **Figure 4.14** we have gene expression evidence that clusters 1,3,4,6, and 7 originate from donor 2 due to expression of non-γδ-T cell markers (CD4-T cells, CD8-T cells and T regulatory cells).

Additionally, the expression of pan-γδ-T markers can confirm the above

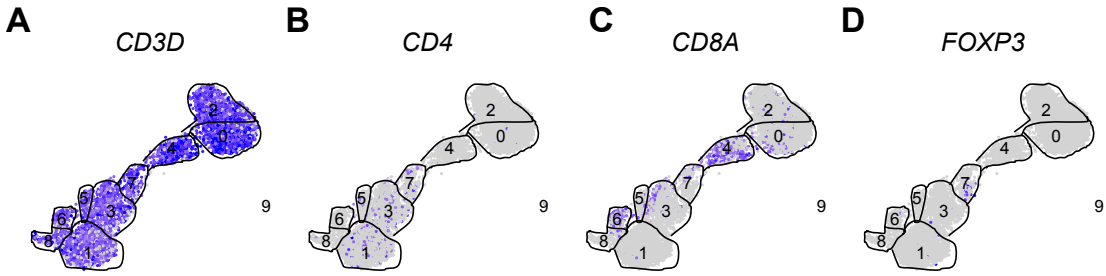



**Figure 4.14:** Expression levels of known αβ-T cell markers in the clusters of the pooled dataset of CD3+ T and γδ-T cells. Purple indicates high expression and grey indicates low expression. **(A)** Expression levels of *CD3E* confirms that all cells are T cells. **(B)** Expression levels of *CD4* indicates that clusters 1, 3 and 7 are possibly CD4-T cells, thus originating from donor 2. **(C)** Expression levels of *CD8A* indicates that clusters 4 and 6 are possibly CD8-T cells, thus originating from donor 2. **(D)** Expression of *FOXP3* in combination with expression of *CD4* in cluster 7 indicates that this cluster is T regulatory cells, further corroborating that cluster 7 cells originate from donor 2.

conclusions and identify the γδ-T cell clusters. Based on the expression of *TRDC* (**Figure 4.15 A**), *TRDV2* (**Figure 4.15 B**) and *TRGV9* (**Figure 4.15 C**), there is strong evidence for clusters 0 and 2 being δ2 γδ-T cells and thus originating from donor 1. Evidence of *TRDV2* and *TRGV9* expression in cluster 5 (**Figure 4.15 B,C**) additionally indicates these are also δ2 γδ-T cells, according to our conclusions in Chapter 3. Thus, cluster 5 also originates from donor 1. Similarly, there is some evidence that cluster 8 consists of δ1 γδ-T cells from donor 1, and high sparsity could be due to higher dropout rates of these genes as observed in Chapter 3.

Finally, cluster 9 cells do not have evidence of expression of any of the above markers. Additionally, the distance from the other clusters on the UMAP could indicate that it is contamination from one of the samples. Thus, these cells were not considered in the comparison. The final dataset consists of 6674 cells, 3102 cells from donor 1 (clusters 0, 2, 5 and 8) and 3572 cells from donor 2 (clusters 1, 3, 4, 6, and 7).

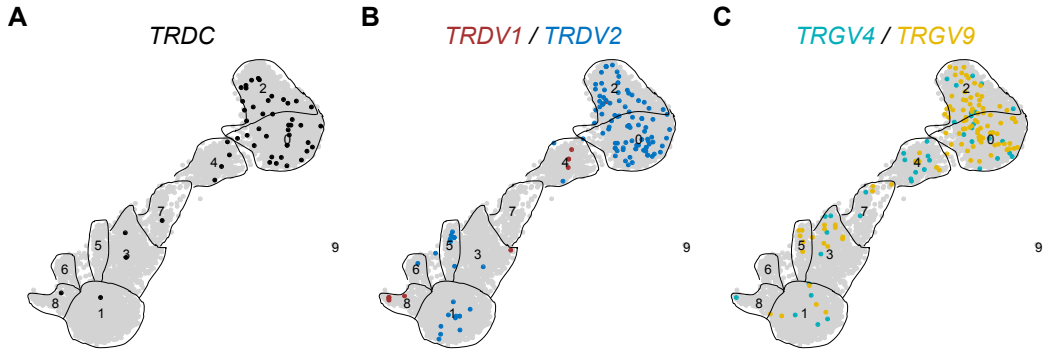



**Figure 4.15:** Manual identification of γδ-T cells based on expression of known γδ-T cell markers. **(A)** Projection of *TRDC+* (black) and *TRDC-* (grey) cells on the UMAP. **(B)** Projection of *TRDV1+* (red) and *TRDV2+* (blue) cells on the UMAP. Grey represents cells that do not express any of these two markers. **(C)** Projection of *TRGV4+* (light blue) and *TRGV9+* (yellow) cells on the UMAP. Grey represents cells that do not express any of these two markers.

Using the sequence alignment file of this dataset, 127,742 SNPs were identified in these cells. The matrix of SNPs by cells is even sparser than the gene expression matrix with 40.52% of the SNPs being only present in a single cell and the median number of SNPs per cell being 498 (**Figure 4.16**).

### Effect of SNP filtering in classification accuracy

To reduce the dimensions of the variants matrix and remove noisy SNPs, I filtered SNPs that are present in very few cells. Again, based on the distribution of SNPs' frequency (**Figure 4.16 B**), there is no obvious threshold to select. Thus, I tried different thresholds of minimum number of cells per SNP from the set of $\{1, 2, 3, 4, 5, 10, 15, 25, 30, 40, 50\}$. **Figure 4.17** shows the effect of SNP filtering in the classification accuracy for different combinations of hyperparameters. The classification accuracy as measured by the Adjusted Rand Index (ARI) between the "ground truth" and the predicted labels is shown on the $y$-axis. Each boxplot shows the ARI for a specific threshold of SNPs' prevalence (shown on $x$-axis) and

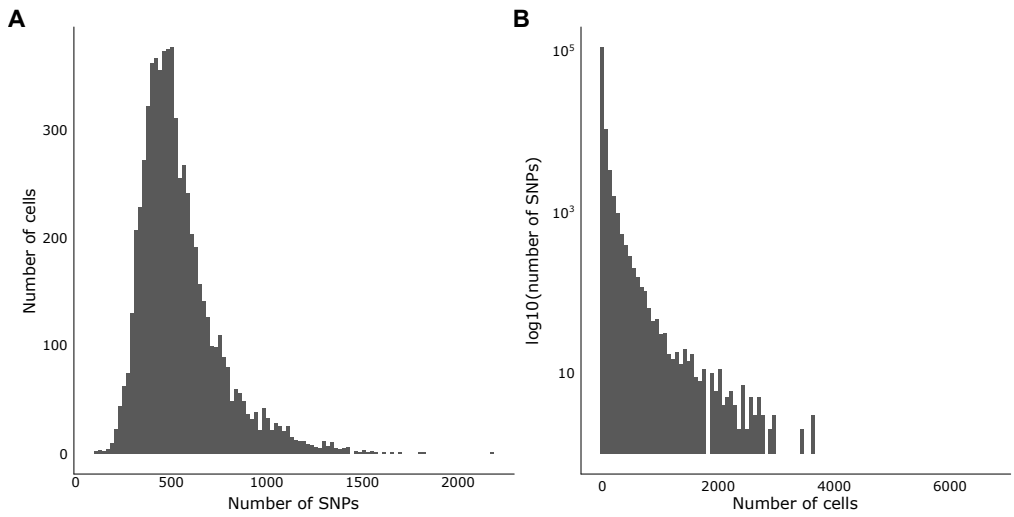



**Figure 4.16:** Distribution of the identified SNPs in the pooled dataset of donors 1 and 2, **(A)** Histogram showing number of identified SNPs per cell. **(B)** Histogram showing frequency of the identified SNPs. $x$-axis shows number of cells a SNP is present and $y$-axis shows the number of SNPs in $log_{10}$ scale. 40.52% of the SNPs are only present in a single cell. SNPs, single nucleotide polymorphisms

for different values of the previously mentioned hyperparameters (dots). Although the hyperparameters have a great effect on the classification accuracy, we can see that the average ARI is improving after filtering SNPs that are present in fewer than 10 cells.

I further tested whether there are any systematic differences in the ARI between the different filtering thresholds using multiple Wilcoxon signed rank tests. The results are summarised in **Table 4.2**, which shows the significance levels for each pair of filtering thresholds, with "*" indicating $p - value \leq 0.01$, "**" indicating $p-value \leq 0.001$ and "***" indicating $p-value \leq 0.001$ and black cells indicating non-significant differences. Only the upper triangle of the table is calculated as the comparisons are symmetrical. Missing comparisons are shown in grey colour. Based on these results, a threshold of 40, results in significantly higher

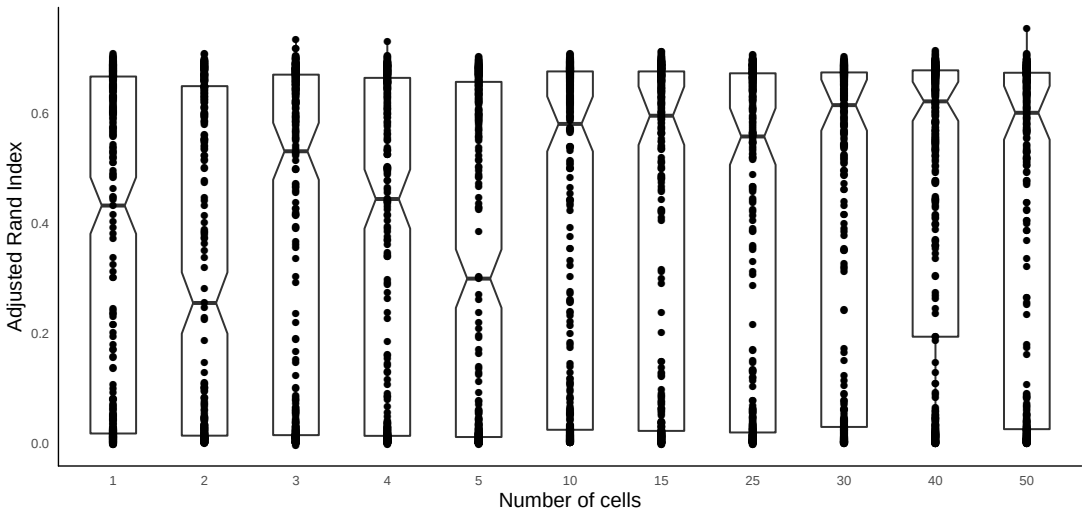



**Figure 4.17:** Boxplot showing classification accuracy as measured by the Adjusted Rand Index ($y$-axis) for different filtering thresholds of SNPs based on their prevalence ($x$-axis). Each dot represents the Adjusted Rand Index for the threshold indicated on the $x$-axis and a different set of hyperparameters.

ARI compared to most of the other thresholds, and thus will be used for the rest of the analysis.

### Effect of depth of network

I next tested the accuracy of PAE_NMF models with different numbers of nodes ranging between 10 and 300. **Figure 4.18** shows the ARI of the classification of these models for different numbers of nodes in layer 1 (grey) and layer 2 (yellow). Each dot represents the average ARI of multiple models with the same number of nodes (indicated on the $x$-axis) and different values of the other hyperparameters (number of epochs and batch size) and error bars indicate 95% confidence interval. Wilcoxon signed rank test was used to test if there are significant differences in classification accuracy between the various selected numbers of nodes. Although there is a trend of decreasing ARI with increasing number of nodes in layer 1, there is no constant significant difference between all pairs of values. For example,

**Table 4.2:** Summary table of pairwise Wilcoxon signed rank tests between ARI of different filtering thresholds (rows and columns). Only the upper triangle of the table is calculated as the comparisons are symmetrical. Missing comparisons are shown in grey colour. "*" indicates $p - value \leq 0.01$, "**" indicates $p - value \leq 0.001$ and "***" indicates $p - value \leq 0.001$ and black cells indicates non-significant differences ($p - value > 0.01$).

| | **1** | **2** | **3** | **4** | **5** | **10** | **15** | **25** | **30** | **40** | **50** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | | | | | | | * | | *** | *** | ** |
| **2** | | | | | | ** | *** | * | *** | *** | *** |
| **3** | | | | | | | | | * | *** | |
| **4** | | | | | | | * | | ** | *** | * |
| **5** | | | | | | * | ** | * | *** | *** | *** |
| **10** | | | | | | | | | | ** | |
| **15** | | | | | | | | | | | |
| **25** | | | | | | | | | | *** | |
| **30** | | | | | | | | | | | |
| **40** | | | | | | | | | | | * |
| **50** | | | | | | | | | | | |

although using 10 nodes is better than using 50 or 100 nodes, it is not better than using 150 nodes. On the contrary, ARI is increasing with increasing number of nodes in layer 2. However, again this difference is not significant between all pairs of values, thus there is no confidence that this can be a generalised rule for designing future models.

Based on these observations, we do not have the power to decide which option is the most appropriate. To test the rest of the hyperparameters, I will be using various models with different numbers of nodes.

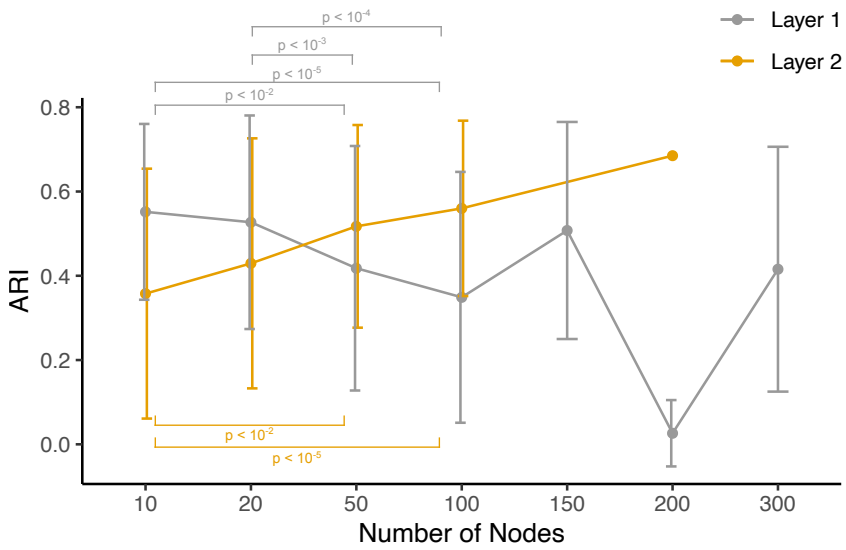



**Figure 4.18:** Effect of number of nodes of the hidden layers in the classification accuracy. Scatter plot showing classification accuracy as measured by the Adjusted Rand Index ($y$-axis) for different number of nodes ($x$-axis) of layer 1 (grey) and layer 2 (yellow). Dots show average ARI of multiple models with the same number of nodes and different values of the other hyperparameters (number of epochs and batch size) and error bars indicate 95% confidence interval. Wilcoxon sign rank test was used to compare the accuracy between pairs of selected numbers of nodes and only significant values are shown. In all tests, the SNP filtering threshold is fixed at 40.

### Effect of number of epochs in the classification accuracy

As expected, an increased number of epochs leads to increased classification accuracy. **Figure 4.19** shows the classification accuracy of the network with different hyperparameters and SNP filtering threshold fixed at 40. The ARI is improving with increased number of epochs up to 30 epochs. However, after 30 epochs, any further increase does not lead to significantly better ARI. For the evaluation of the remaining hyperparameters I only regarded results with 30-100 epochs.

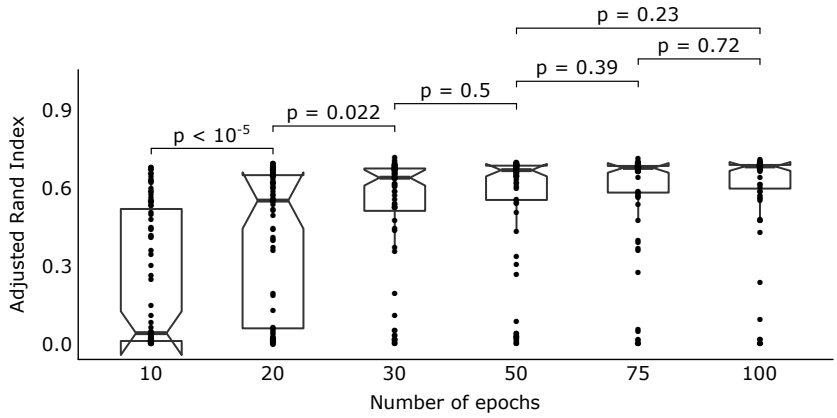



**Figure 4.19:** Effect of number of epochs on the classification accuracy. Boxplot showing classification accuracy as measured by the Adjusted Rand Index ($y$-axis) for different numbers of epochs ($x$-axis). Each dot represents the Adjusted Rand Index for the number of epochs indicated on the $x$-axis and a different set of hyperparameters. In all tests, the SNP filtering threshold is fixed at 40.

### Effect of batch size in the classification accuracy

Here I tested the accuracy of classification for batch sizes ranging between 4 and 512 (**Figure 4.20**) due to the increased number of cells compared to the previous example. A batch size of 16 seems to achieve the best accuracy for a set of tests with different hyperparameters and with SNP filtering threshold of 40. Pairwise Wilcoxon signed rank tests between results with batch size 16 and results with

other batch sizes confirms that classification accuracy is significantly higher when batch size is set to 16 compared to any other value.

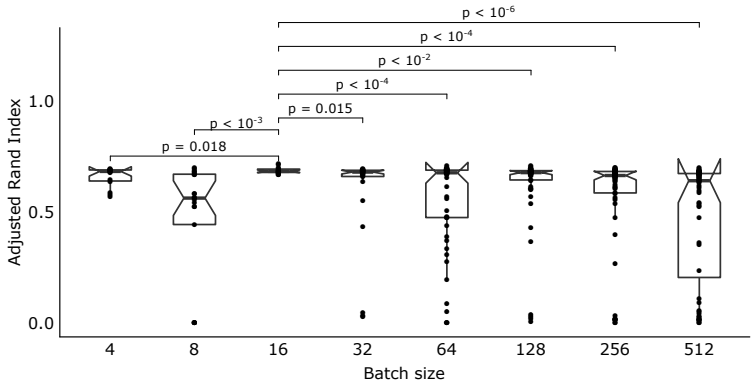



**Figure 4.20:** Effect of batch size in the classification accuracy. Boxplot showing classification accuracy as measured by the Adjusted Rand Index ($y$-axis) for different batch sizes ($x$-axis). Each dot represents the Adjusted Rand Index for the batch size indicated on the $x$-axis and a different set of hyperparameters. P-values indicate significance level of ARI with batch size 16 being higher than the ARI of any other group using an one-sided Wilcoxon signed rank test. In all tests, the SNP filtering threshold is fixed at 40 and epochs are between 30 and 100.

### Effect of batch size in loss

As seen in the previous dataset, the loss is highly dependent on the batch size (**Figure 4.21**) with higher batch size leading to lower loss. This however does not correspond to better classification accuracy (**Figure 4.20**). For this reason, we will only compare different PAE_NMF models with the same batch size to select the one with the lowest loss for the final classification.

### Extraction of donor-specific variants for validation and doublet detection

From the W matrix, I extracted the top 10 genetic variants that contribute most to the identity of each donor. **Figure 4.22 A** shows the extracted variants sorted

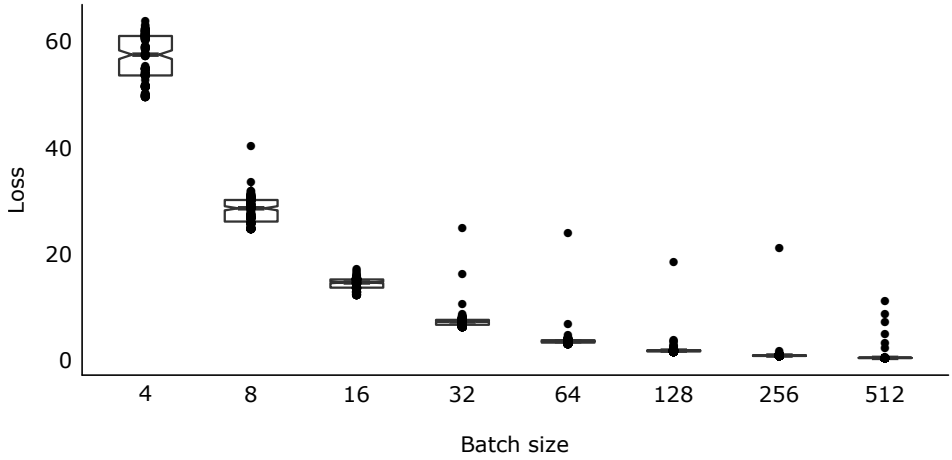



**Figure 4.21:** Boxplot showing loss calculated after the last epoch for different batch sizes. Each dot represents the loss ($y$-axis) of a model with batch size indicated on the $x$-axis and a different set of hyperparameters (number of epochs, numbers of nodes, SNP filtering threshold).

by their weight. The length of the bars represents the variants' weight with positive values indicating enrichment in donor 1 and negative values indicating enrichment in donor 2. Variants are encoded as "chromosome_position". **Figure 4.22 B** heatmap shows the presence of these selected variants (rows) in all cells (columns), grouped by donor identity according to the PAE_NMF method. Red represents presence and blue represents absence of a variant in a cell. Due to the within-donor heterogeneity, there is higher variability within each donor and more noise compared to the previous example. However, there are at least 3 variants unique to one of the donors and captured in all donor's cells.

### Doublet detection

Using hierarchical clustering with Euclidean distance to group the cells into three clusters, we obtain the refined classification shown in **Figure 4.23**. 2914 cells are assigned to donor 1, 3729 cells are assigned to donor 2 and 180 cells are identified as doublets.

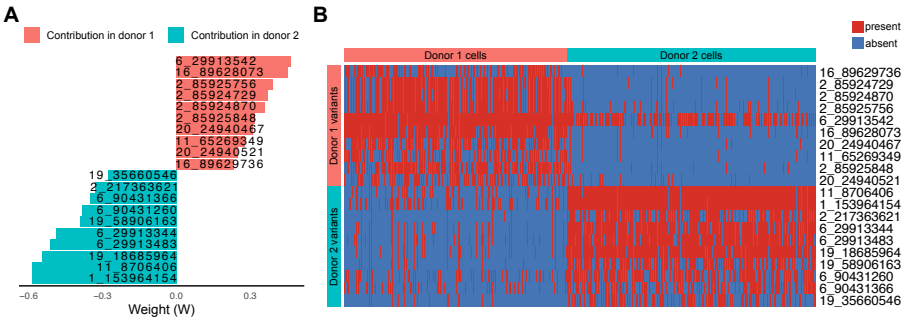



**Figure 4.22:** Extraction of donor-specific variants. **(A)** Barplot showing top 10 variants per donor sorted by weight. The length of the bars represents the weight with positive values indicating enrichment in donor 1 and negative values indicating enrichment in donor 2. Variants are encoded as "chromosome_position". **(B)** Heatmap showing the presence (red) or absence (blue) of the top 10 variants per donor (rows) in cells (columns) grouped by donor identity according to the PAE_NMF method.

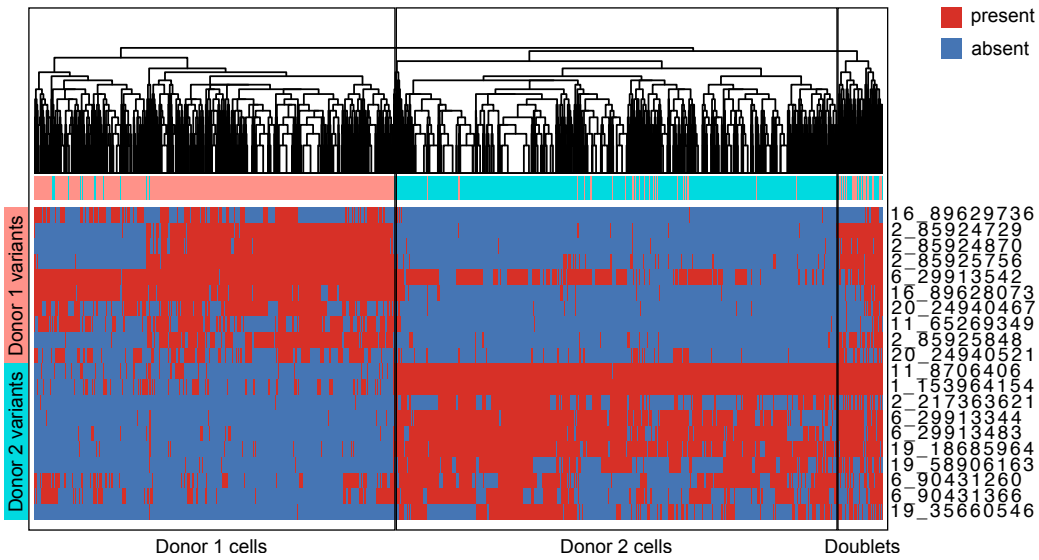



**Figure 4.23:** Doublet detection with hierarchical clustering. Heatmap showing the presence (red) or absence (blue) of the top 10 variants per donor (rows) in cells (columns) grouped by hierarchical clustering using Euclidean distance. Each identified group of cells is manually annotated.

### Comparison with existing methods

Next, I compared the accuracy of the classification of PAE_NMF with and without identification of doublets with that of vireo and scSplit. Demuxlet and cardelino

cannot be used with this data due to unavailable genotypic information of the two donors.

**Figure 4.24 A** shows the ARI between the predicted labels from each method and the "ground truth" labels. Since the "ground truth" labels do not include a class for doublets, predicted doublets from each method have been removed from the comparison. For PAE_NMF, eight different results have been used, the ones with the lowest loss between all models with equal batch size and the error bar shows the 95% confidence interval. The initial result of PAE_NFM has lower ARI than the other methods. However, the refined labels after hierarchical clustering with the top 10 variants per donor shows considerably improved accuracy, very similar to that of vireo. scSplit has the highest ARI (equal to 1), however, almost 40% of the cells are identified as doublets indicating low sensitivity, as doublets are not expected to exceed 10% of the cells. Indeed, the percentage of identified doublets by vireo and PAE_NMF with hierarchical clustering are below 5% which is much closer to our expectation. Finally, 90% of the identified doublets with PAE_NMF (163 cells) have been also identified as doublets by one or both of the other methods (**Figure 4.24 C**).

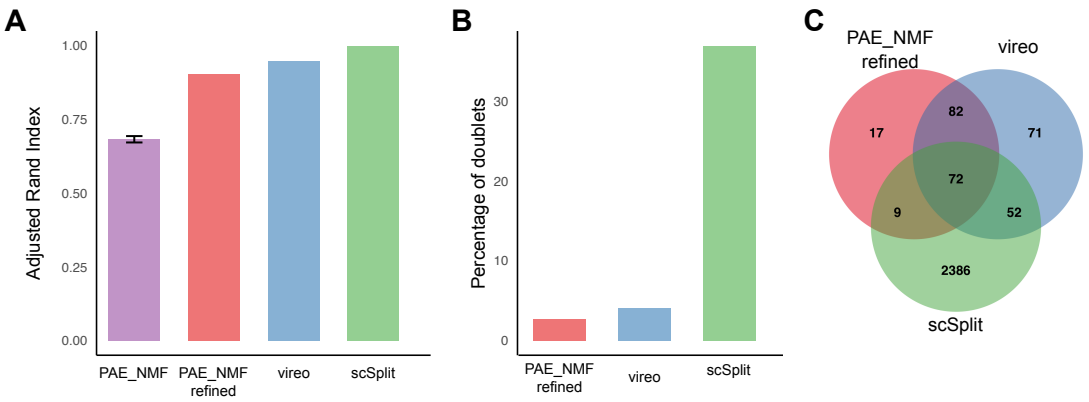



**Figure 4.24:** Comparison of PAE_NMF to vireo and cardelino. **(A)** Barplot showing Adjusted Rand Index ($y$-axis) between the predicted labels from each method indicated on the $x$-axis and the "ground truth" labels. Doublets have been removed from the comparison since the "ground truth" labels do not contain any doublet class. For PAE_NMF, eight different results have been used, the ones with the lowest loss between all tests with equal batch size and the error bar shows the 95% confidence interval. "PAE_NMF_refined" is the result from the combination of PAE_NMF and hierarchical clustering using the top 10 variants per donor. **(B)** Barplot showing percentage of identified doublets from each method. **(C)** Venn diagram showing overlap of identified doublets between the three methods.

## 4.4 Discussion

In this chapter, I presented a deep learning approach for demultiplexing the donor identities of cells in pooled single-cell RNA-seq experiments. An implementation of non-negative matrix factorisation using variational autoencoders has two main advantages for the analysis of noisy single-cell RNA-sequencing data. Variational autoencoders aim to only approximate the input instead of perfectly reconstructing it, enabling the network to learn useful properties of the data while ignoring signal noise. Additionally, the use of non-negative matrix factorisation enables tracking of the genetic variants that contribute to the definition of each donor. These variants can then be used for validation of the clustering. Using two different datasets, I explored the effect of the hyperparameters of the model in the classification accuracy and compared it to other published methods, namely demuxlet, vireo and scSplit.

The high dimensionality of the data is the main bottleneck of the data. As previously discussed, besides sparsity and dropouts that also exist in the single-cell gene expression datasets, there are systematic differences between cells of a single donor based on cell type specific gene expression. Filtering very rare SNPs can potentially improve the accuracy of the model by removing too noisy variables without sufficient numbers of measurements. This can also have a positive effect on the computational time of the model. While there is no obvious threshold based on the distribution of the SNPs prevalence, I tried several thresholds and showed that pre-filtering can improve the classification accuracy in the complex dataset of pooled CD3+ T and γδ-T cells that consists of a very high number of identified SNPs.

An alternative filtering approach could be to select only variants in the genomic regions of housekeeping genes, in order to ensure there is equal probability of

capturing a SNP in all donor's cells. However, this runs the risk of reducing the number of observed variants and might lead to increasing the number of cells with no SNPs. Additionally, the set of housekeeping genes is relative to the cell populations in the mixture and it is challenging to identify them within each dataset separately with minimum user interaction and expertise.

Besides the filtering threshold, it is known that the performance of variation autoencoders is strongly dependent on the tuning of the hyperparameters (Hu and Greene (2019), Eraslan et al. (2019)). In the results presented here we observe very high variation of the classification accuracy depending on the hyperparameters used. As expected, increased number of epochs significantly increase the classification accuracy. Regarding the depth of the network, i.e. the number of nodes in each hidden layer, the results do not show a clear dependence and a rule that can be generalised for the design of new models.

The final reconstruction loss is expected to correlate with the classification accuracy, with low loss indicating better hyperparameter configuration and better accuracy Eraslan et al. (2019). Thus, in a similar method for denoising and imputation of single-cell RNA-sequencing data (DCA), Eraslan et al. (2019) are implementing a hyperparameter search. The DCA is trained with 1,000 different hyperparameter configurations, and the model with the lowest reconstruction error is selected as the final result.

Here we have found that the loss is indicative of the classification accuracy when comparing models with the same batch size. However, the loss is highly dependent on the batch size selected which does not correlate with the classification accuracy. The solution that worked for both datasets was to train multiple models with equal batch size and different sets of the rest of the hyperparameters and select the model with the lowest loss but not compare across models with different batch sizes. This assumption needs to be confirmed for a higher number of

datasets. Given the low computational time and the ability to train the different models in parallel, this is a feasible solution without requirement of user-specified hyperparameters.

Identification of doublets has not been implemented in the PAE_NMF method. One suggestion for improvement could be to increase the number of requested clusters in order to account for extra clusters consisting of doublets. This was not successful with the above datasets, possibly due to the fact that the doublets do not have unique SNPs, but a combination of SNPs observed in two different donors. To account for this, here I extended the method by adding a last step of hierarchical clustering of the cells using the top 10 most important SNPs selected from the **W** matrix. This was shown to improve the classification accuracy of PAE_NMF and there is high confidence that the cells detected as doublets are correct based on the agreement with the other methods. It is however unclear whether this can be generalised in datasets with a very high number of donors.

Here I have only used top 10 SNPs per donor to refine the clustering of the PAE_NMF model and detect doublets. Further exploration of the resulting **W** matrix is required to develop a generalisable approach for identifying all SNPs that significantly contribute to a donor's identity and test whether this can further improve the classification accuracy.

Overall, the classification accuracy of the selected models with minimum loss was close to that of vireo and demuxlet after refinement of the cell labels with hierarchical clustering. While scSplit seems to correctly label all cells identified as non-doublets showing a very high specificity, the sensitivity is very low with a very high number of cells labelled as doublets.

This study was restricted to testing the model with two datasets due to lack of publicly available datasets with ground truth labels of the donor identities. While

it was shown that PAE_NMF can overcome the bias of intra-donor variability due to celltype-specific gene expression, this was only demonstrated for a dataset (PBMCs) comprising cells of two different donors. To further test the classification accuracy of this method and it's potential for contribution in population-scale studies, more complex datasets are required that comprise of cells of a higher number of donors and cells.

### 4.4.1 Conclusions

There has been an increased interest recently in using deep learning methods for the analysis of single-cell RNA-sequencing data, with applications on normalisation, clustering and identification of differentially expressed genes. The method presented here is the first application of deep learning on demultiplexing donor identities of cells in pooled experiments. Preliminary results show that although the accuracy is highly dependent on the hyperparameters of the model, it can achieve high classification accuracy with proper hyperparameters. Although no generalised rules could be found to tune the hyperparameters, the reconstruction loss could be used to find the most accurate result amongst models with different settings but same batch size.

The interpretability of non-negative matrix factorisation allows not only the clustering of cells by donor identity, but also the identification of donor-specific genetic variants that distinguish between individuals. This information can be leveraged in other applications. For example, the identified donor-specific variants can be removed from the sequencing data prior to them becoming publicly available in order to become anonymised. Finally, this approach can be used to other types of highly sparse single-cell data, such as DNA methylation and splicing data.

# Chapter 5

# Discussion

With rapid improvement of sequencing protocols and decreasing sequencing costs, single-cell RNA sequencing has become a routine method for studying cell biology. scRNA-seq enables measurement of gene expression at the individual cell level leading to an understanding of cell-to-cell heterogeneity at greater resolution compared to bulk RNA-sequencing. This new information has helped uncover previously unknown cell types and states.

However, new computational challenges have emerged. The low amounts of mRNA available and amplification biases lead to uncertainty about measured gene expression levels. Relative gene expressions between cells are unbalanced, with no corresponding reads observed for a high number of genes (dropout events) (Kharchenko et al., 2014). Additional systematic technical variability exists between data that have been processed separately, known as a batch effect, that confounds biological variation of gene expression.

This thesis contributes computational methods for the analysis of single-cell RNA-seq data to overcome the bias of batch effects, as well as a better understanding of

the transcriptional and functional heterogeneity of γδ-T cells in human peripheral blood and breast tumour.

**Identification of transcriptionally equivalent cell types across datasets with batch effect**

Cross-sample comparison is essential in order to understand the role and altered states of cell populations in development and disease. Batch effect, however, is posing technical challenges for the integrated analysis of datasets that have been processed separately. While batch effect correction methods exist (Haghverdi et al. (2018), Hoffman et al. (2018)) that can identify and regress out technical variability, these do not work as expected in pairs of data with strong imbalances in numbers of cells and sequencing depth. Additionally, they run the risk of over-correcting and removing true biological variability and finally, they require a repeat of the analysis for any new dataset obtained. A more promising strategy is mapping cells to reference datasets (Lähnemann et al., 2020).

In **Chapter 2** of this thesis I presented a mapping method for the identification of transcriptionally equivalent cell populations across datasets with batch effect. scID uses the framework of Fisher's LDA to score the cells of a dataset for a given gene set and then partitions the cells based on this score as matching and non-matching to that gene set. The gene set is expected to represent a cell type or state and can be either extracted from a reference single-cell RNA-sequencing dataset or given as a list from the user, as a result of curation or analysis of bulk RNA-seq data.

Through extensive evaluation and comparison with other methods, I showed that scID outperforms existing methods in cases of pairs of datasets (reference and target) that have strong imbalances in numbers of cells and sequencing depths. This is due to scID sorting genes based on their discriminative power in the

target dataset which leads to filtering out genes that in the target dataset are not discriminative due to higher dropout rate or different cell type composition. Moreover, gene sets are only extracted once from the reference dataset and can be used with any new target dataset acquired.

An important limitation of scID is that a cell might be selected by more than one gene set if these represent transcriptionally similar cell types. The current implementation resolves these multi-assignments by comparing the scores of a cell for all tested gene sets and assigning the cell to the identity with the highest score. However, this temporary solution needs to be replaced by a better approach. Replacing the score with the probability of a cell being drawn from a population described by the given gene set would allow better classification of the cells as well as leave ambiguous cells unclassified to reduce false positives. Inference of minimum probability or score from the reference data could also improve the accuracy of scID but needs further exploration.

**Uncovering previously unappreciated heterogeneity with the γδ-T cell population**

While γδ-T cells in PBMC have been previously characterised based on the variable segment of the delta chain of the TCR (Pizzolato et al., 2019) or based on production of IL17 or INFγ (Ribot et al., 2009), the underlying heterogeneity of these subtypes was not understood. In **Chapter 3** using scID and other published scRNA-seq data analysis methods, I sought to unbiasedly identify transcriptionally and functionally distinct subpopulations of γδ-T cells in human peripheral blood (PBMC) and breast tumour samples.

Unsupervised clustering revealed five subpopulations of γδ-T cells in PBMC and three subpopulations of γδ-T cells in breast tumour samples. Two of the γδ-T cell subtypes in PBMC are possibly δ1 and three are δ2 which confirms what is known

regarding the relative prevalence of δ1 and δ2 subsets in human peripheral blood. There is evidence that one of the γδ-T cell subtypes in breast tumour samples is δ2 while the TCR δ chain identity of the other two clusters is unclear. Based on gene markers and comparison of gene expression signatures with the clusters in PBMC, we inferred that one of the clusters is most like δ1 and the other δ2. While in these experiments the 3' UTR chemistry of Chromium 10X was used, the 5' UTR chemistry could capture the δ and γ variable region of each cell with higher confidence and lower dropout rate. Thus, this preliminary characterisation of the subpopulations needs further validation.

Regardless of the γ and δ chains, the identified subtypes showed different gene expression patterns and annotated functions. Extended lists of cluster-specific genes and annotated functions are provided for further study of these subpopulations either through sorting with cell surface markers or computationally using a longer set of genes. Interestingly, a breast tumour subtype was associated with improved overall survival of breast cancer patients. While some well-known factors associated with survival, such as the mutation load, the overall levels of T-cells and the expression of NKG2D ligands, could be eliminated, other factors could be confounding the result. For example, since NK cells are absent from the two breast cancer datasets, it is possible that the extracted gene set is also expressed in NK cells and thus the high scoring samples from TCGA are selected due to higher proportion of NK cells instead of this γδ-T cell subtype. Refinement of the selected gene set is required to eliminate this factor. Additionally, it would be interesting to investigate whether enriched samples from other cancer studies and types yield similar results.

It has been previously seen that there is variation in the relative proportion of δ2 subtypes between different individuals (Ryan et al., 2016). The new data presented in this thesis could be also used to observe such inter-donor variation of the identified subtypes. To achieve this, we can use the PAE_NMF method

presented in **Chapter 4** to demultiplex the donor identities of the cells of the pooled HD4/5 dataset in order to quantify the relative proportion of all cell subtypes in each of the three PBMC donors.

### Demultiplexing donor identity of cells in pooled scRNA-seq experiments

As ultra high throughput single-cell RNA-sequencing becomes efficient and cost effective (Zhang et al., 2019a), another way to overcome batch effects is to pool cells from multiple individuals in one experiment. In cases where donors represent different conditions, e.g. different developmental stages or health conditions, such datasets enable differential gene expression analysis between conditions with existing methods. Additionally, it can scale single-cell RNA-seq data analysis to the population level, due to a decrease of the per-individual library cost. Using data from multiple donors in a study increases the confidence and reproducibility of the results, while enabling the comparison of cell type composition between individuals.

Such an approach requires tracking of cells' donor identity. Experimental methods that enable tracking of donor identity of each cell, such as SPLiT-Seq (Rosenberg et al., 2018) and CITE-Seq (Stoeckius et al., 2018), require heavy manual processing and are costly. Computationally, donor identities can be assigned to cells based on donor-specific genetic variants. Technical factors, such as sparsity and gene fragment capturing, as well as biological factors, such as cell-type-specific gene expression, need to be addressed in such computational methods.

In **Chapter 4** I presented a deep learning application for non-negative matrix factorisation using Variational Autoencoders (VAEs) to identify the donor identity of cells by clustering them using the observed genetic variants. NMF is a factorisation of the input genetic variants matrix into two non-negative matrices,

with one representing the load of each variant in each donor's identity and the other representing the probability of cell belonging to each donor. VAEs add a probabilistic extension to NMF that works as denoising. The input is approximated instead of perfectly reconstructed which leads to retention of only informative features of the data without the risk of overfitting. This enables both clustering the cells based on their donor identity, as well as extracting donor-specific variants that can be used for validation.

Although the accuracy of this method is highly dependent on the selection of hyperparameters of the model, the final reconstruction loss can be used to compare models with different sets of hyperparameters. The model with the lowest loss can be selected and is expected to have good performance. Since this was not extensively evaluated, further testing is required to prove the validity of this hypothesis. Finally, it would be of great interest to test the method for datasets with more than two donors. All other methods have shown good classification accuracy for datasets with up to 8 donors, while vireo shows decreased performance for synthetic datasets with more than 12 donors (Huang et al., 2019). PAE_NMF needs to be tested with such complex datasets to investigate its potential advantage over existing methods and ensure that it can be used at population-scale scRNA-seq studies.

## 5.1 Open challenges in single-cell RNA-sequencing data analysis

Even though many methods for the analysis of single cell RNA-seq data, such as dimensionality reduction, clustering and visualisation, are becoming parts of standardised pipelines, there are still many open computational challenges.

### Differential gene expression

For the improvement of scID, the most important open challenge is differential gene expression analysis for the extraction of cluster-specific genes from reference scRNA-seq datasets. Identification of cluster-specific gene sets needs to be separated into two problems, depending on the scope of the study.

When extracted genes are expected to be used as biomarkers for sorting a cell population, selected genes need to be either exclusively present or the distributions of gene expressions in that population and in any other cell should not be highly overlapping. Markers often need to be hierarchical to provide a set of biomarkers that need to be combined in order to distinguish between similar subpopulations. An example of a hierarchical approach for defining gene expression profiles of cell types is CHETAH (de Kanter et al., 2019).

On the other hand, for any computational analysis, such as functional annotation or mapping across datasets, longer lists of genes that in combination are cluster-specific need to be identified. As discussed in Chapter 1, current methods for differential gene expression analysis, such as `MAST` (Finak et al., 2015) used in scID, identify differentially expressed genes in a cluster by testing for differences in the gene expression or the distribution of the gene expression between that cluster and a pool of cells from the other clusters in the dataset. This leads to identifying genes as differentially expressed those that are also expressed in other similar cell populations that do not contribute equally to the signal of the pool of cells, given that they are a minor population.

A solution for this problem within scID could be to perform multiple pairwise comparisons between clusters to find a set of genes for a reference cluster $c$ that can distinguish between cells of type $c$ and cells of transcriptionally close clusters. Using the minimax approach (Devroye and Lugosi, 2001), we can find a

feature vector $\mathbf{w}_c = [w_1, ..., w_M]_c^T$ for each reference cluster $c$ of $M$ genes and their corresponding weights that define a one-dimensional projection of the reference cells so that cluster $c$ is maximally separated from the transcriptionally closest cluster, and thus from any cluster in the reference data. This can be achieved by solving the following problem

$$\arg\max_{w} \min_{j} \frac{(\mathbf{w}^T(\boldsymbol{\mu}_c - \boldsymbol{\mu}_j))^2}{\mathbf{w}^T(\Sigma_c + \Sigma_j)\mathbf{w}} \tag{5.1}$$

where $\boldsymbol{\mu}$ is a vector of the average expression of each of the $M$ genes in cluster $c$ ($\boldsymbol{\mu}_c$) or any other cluster $j$ ($\boldsymbol{\mu}_j$) and $\Sigma$ is the covariance matrix between genes in cluster $c$ ($\Sigma_c$) and any other cluster $j$ ($\Sigma_j$).

Unlike the currently implemented method of Step 2 of scID, in this approach the cost is computed in the projected line instead of each gene individually. Preliminary results showed better separation between clusters of various reference datasets and did include known markers for each cell type. However, the need to calculate the full covariance as well as perform all these pairwise comparisons is computationally intensive and did not scale to greater than 3000 genes. Restriction of the number of genes does not only require supervised gene filtering prior to applying this method, but can also lead to decreased accuracy in cases with many and transcriptionally similar clusters in the data, which requires longer lists of features to distinguish them. Working towards the direction of improving the computational time of the method might enable the implementation of this method within scID but also as an alternative method for differential gene expression analysis.

Finally, another limitation of the cluster-specific gene sets extracted from the reference clusters in scID, is that these gene sets are relative to the other cell types present in the reference dataset. This does not guarantee that a cell population in

a target dataset that is closely related but not identical to the reference population will not be selected as matching. There is thus the need to build better reference gene sets that consist of genes differentially expressed between the cell population of interest and any other population that can be possibly present in the same type of tissue. Datasets from the Human Cell Atlas can serve as an important reference resource. The main challenge of this is the presence of a batch effect between datasets, thus differential gene expression analysis methods for integrated data are required.

### Data integration and scaling

As protocols improve and the cost of single-cell RNA-sequencing decreases, bigger and more heterogeneous types of single-cell data can be generated for a single study. This will need Big Data methods for the analysis that will involve both the ability of methods to scale up to more cells as well as to combine different types of measurements into the same analysis.

In addition to single-cell RNA-sequencing, methods exist for measuring DNA methylation (Karemaker and Vermeulen, 2018), protein levels (Virant-Klun et al., 2016) as well as chromatin accessibility (Cusanovich et al., 2015) at the single cell level. Integration of genomic, proteomic and epigenomic information can lead to better resolution of cell types and their distinct functions, however, at the cost of further increasing the dimensionality of the data. The increase of both the number of features as well as the number of cells per dataset will be challenging for many in-memory processing methods and interest should be shifted to parallel processing. Finally, an important challenge of such data integration is to model the dependencies between the different measurements.

# References


Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 1973, pages 420–434. Springer Verlag.

Altvater, B., Pscherer, S., Landmeier, S., Kailayangiri, S., Savoldo, B., Juergens, H., and Rossig, C. (2012). Activated human $\gamma\delta$ T cells induce peptide-specific CD8+ T-cell responses to tumor-associated self-antigens. *Cancer Immunology, Immunotherapy*, 61(3):385–396.

Andrews, T. S. and Hemberg, M. (2019). M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics (Oxford, England)*, 35(16):2865–2867.

Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., Choi, K., Fromme, R. M., Dao, P., McKenney, P. T., Wasti, R. C., Kadaveru, K., Mazutis, L., Rudensky, A. Y., and Pe'er, D. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, 174(5):1293–1308.e36.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17(1):63.

Bauer, S., Groh, V., Wu, J., Steinle, A., Phillips, J. H., Lanier, L. L., and Spies, T. (1999). Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA. *Science*, 285(5428):727–729.

Bayar, B., Bouaynaya, N., and Shterenberg, R. (2014). Probabilistic non-negative matrix factorization: Theory and application to microarray data analysis. *Journal of Bioinformatics and Computational Biology*, 12(1).

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–47.

Benevides, L., Da Fonseca, D. M., Donate, P. B., Tiezzi, D. G., De Carvalho, D. D., De Andrade, J. M., Martins, G. A., and Silva, J. S. (2015). IL17 promotes mammary tumor progression by changing the behavior of tumor cells and eliciting tumorigenic neutrophils recruitment. *Cancer Research*, 75(18):3788–3799.

Bengtsson, M., Hemberg, M., Rorsman, P., and Ståhlberg, A. (2008). Quantification of mRNA in single cells and modelling of RT-qPCR induced noise. *BMC Molecular Biology*, 9(1):63.

Billadeau, D. D., Upshaw, J. L., Schoon, R. A., Dick, C. J., and Leibson, P. J. (2003). NKG2D-DAP10 triggers human NK cell-mediated killing via a Syk-independent regulatory pathway. *Nature Immunology*, 4(6):557–564.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Brandes, M., Willimann, K., and Moser, B. (2005). Immunology: Professional antigen-presentation function by human $\gamma\delta$ cells. *Science*, 309(5732):264–268.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1098.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, 16(1):43–49.

Carmi, Y., Rinott, G., Dotan, S., Elkabets, M., Rider, P., Voronov, E., and Apte, R. N. (2011). Microenvironment-Derived IL-1 and IL-17 Interact in the Control of Lung Metastasis. *The Journal of Immunology*, 186(6):3462–3471.

Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 785–794. Association for Computing Machinery.

Chung, N. C. and Storey, J. D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M., Factor, R. E., Collins, L. C., Allison, K. H., Chen, Y. Y., Jensen, K., Johnson, N. B., Oesterreich, S., Mills, G. B., Cherniack, A. D., Robertson, G., Benz, C., Sander, C., Laird, P. W., Hoadley, K. A., King, T. A., Akbani, R., Auman, J. T., Balasundaram, M., Balu, S., Barr, T., Benz, S., Berrios, M., Beroukhim, R., Bodenheimer, T., Boice, L., Bootwalla, M. S., Bowen, J., Brooks, D., Chin, L., Cho, J., Chudamani, S., Davidsen, T., Demchok, J. A., Dennison, J. B., Ding, L., Felau, I., Ferguson, M. L., Frazer, S., Gabriel, S. B., Gao, J. J., Gastier-Foster, J. M., Gehlenborg, N., Gerken, M., Getz, G., Gibson, W. J., Hayes, D. N., Heiman, D. I., Holbrook, A., Holt, R. A., Hoyle, A. P., Hu, H., Huang, M., Hutter, C. M., Hwang, E. S., Jefferys, S. R., Jones, S. J., Ju, Z., Kim, J., Lai, P. H., Lawrence, M. S., Leraas, K. M., Lichtenberg, T. M., Lin, P., Ling, S., Liu, J., Liu, W., Lolla, L., Lu, Y., Ma, Y., Maglinte, D. T., Mardis, E., Marks, J., Marra, M. A., McAllister, C., Meng, S., Meyerson, M., Moore, R. A., Mose, L. E., Mungall, A. J., Murray, B. A., Naresh, R., Noble, M. S., Olopade, O., Parker, J. S., Pihl, T., Saksena, G., Schumacher, S. E., Shaw, K. R., Ramirez, N. C., Rathmell, W. K., Roach, J., Robertson, A. G., Schein, J. E., Schultz, N., Sheth, M., Shi, Y., Shih, J., Shelley, C. S., Shriver, C., Simons, J. V., Sofia, H. J., Soloway, M. G., Sougnez, C., Sun, C., Tarnuzzer, R., Tiezzi, D. G., Van Den Berg, D. J., Voet, D., Wan, Y., Wang, Z., Weinstein, J. N., Weisenberger, D. J., Wilson, R., Wise, L., Wiznerowicz, M., Wu, J., Wu, Y., Yang, L., Zack, T. I., Zenklusen, J. C., Zhang, J., Zmuda, E., and Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, 163(2):506–519.

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., and Yosef, N. (2019). Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems*, 8(4):315–328.e8.

Collins, F. S., Lander, E. S., Rogers, J., and Waterson, R. H. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Coons, A. H., Creech, H. J., and Jones, R. N. (1941). Immunological Properties of an Antibody Containing a Fluorescent Group. *Proceedings of the Society for Experimental Biology and Medicine*, 47(2):200–202.

Cragg, J. G. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica*, 39(5):829.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex

single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914.

Davey, M. S., Willcox, C. R., Hunter, S., Kasatskaya, S. A., Remmerswaal, E. B., Salim, M., Mohammed, F., Bemelman, F. J., Chudakov, D. M., Oo, Y. H., and Willcox, B. E. (2018). The human V$\delta$2+ T-cell compartment comprises distinct innate-like V$\gamma$9+ and adaptive V$\gamma$9- subsets. *Nature Communications*, 9(1).

de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T., and Holstege, F. C. (2019). CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic acids research*, 47(16):e95.

demuxlet (2017). demuxlet/tutorial at master · statgen/demuxlet · GitHub. Accessed: 2020-03-23.

Devroye, L. and Lugosi, G. (2001). *Minimax Theory*. Springer, New York, NY.

Dimova, T., Brouwer, M., Gosselin, F., Tassignon, J., Leo, O., Donner, C., Marchant, A., and Vermijlen, D. (2015). Effector V$\gamma$9V$\delta$2 T cells dominate the human fetal gammadelta T-cell repertoire. *Proc Natl Acad Sci U S A*, 112(1091-6490 (Electronic)):E556–E565.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2019). CellPhoneDB v2.0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. *bioRxiv*, page 680926.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1).

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Freeman, W. M., Walker, S. J., and Vrana, K. E. (1999). Quantitative RT-PCR: Pitfalls and potential. 26(1):112–125.

Fulwyler, M. J. (1965). Electronic separation of biological cells by volume. *Science*, 150(3698):910–911.

Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points: Online stochastic gradient for tensor decomposition. In *Journal of Machine Learning Research*, volume 40, pages 1–46.

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., Nair, V. S., Xu, Y., Khuong, A., Hoang, C. D., Diehn, M., West, R. B., Plevritis, S. K., and Alizadeh, A. A. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature Medicine*, 21(8):938–945.

Gierahn, T. M., Wadsworth, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Christopher Love, J., and Shalek, A. K. (2017). Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods*, 14(4):395–398.

Graham, F. L., Smiley, J., Russell, W. C., and Nairn, R. (1977). Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *Journal of General Virology*, 36(1):59–72.

Groh, V., Rhinehart, R., Secrist, H., Bauer, S., Grabstein, K. H., and Spies, T. (1999). Broad tumor-associated expression and recognition by tumor-derived $\gamma\delta$ T cells of MICA and MICB. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6879–6884.

Grønbech, C. H., Vording, M. F., Timshel, P., Sønderby, C. K., Pers, T. H., and Winther, O. (2019). scVAE: Variational auto-encoders for single-cell gene expression data. *bioRxiv*, page 318295.

Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640.

Grün, D., Muraro, M. J., Boisset, J. C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., de Koning, E. J., and van Oudenaarden, A. (2016). De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*, 19(2):266–277.

Haas, J. D., Malinarich González, F. H., Schmitz, S., Chennupati, V., Föhse, L., Kremmer, E., Förster, R., and Prinz, I. (2009). CCR6 and NK1.1 distinguish between IL-17A and IFN-$\gamma$-producing $\gamma\delta$ effector T cells. *European Journal of Immunology*, 39(12):3488–3497.

Habib, N., Avraham-Davidi, I., Hofree, M., Shekhar, K., Gelfand, E., Burks, T., Aguet, F., Ardlie, K., Weitz, D. A., Regev, A., Zhang, F., Choudhury, S. R.,

Habib, N., Basu, A., and Rozenblatt-Rosen, O. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods*, 14(10):955–958.

Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296.

Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G. C., Chen, M., and Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 172(5):1091–1107.e17.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., Zhou, Y., Ye, F., Jiang, M., Wu, J., Xiao, Y., Jia, X., Zhang, T., Ma, X., Zhang, Q., Bai, X., Lai, S., Yu, C., Zhu, L., Lin, R., Gao, Y., Wang, M., Wu, Y., Zhang, J., Zhan, R., Zhu, S., Hu, H., Wang, C., Chen, M., Huang, H., Liang, T., Chen, J., Wang, W., Zhang, D., and Guo, G. (2020). Construction of a human cell landscape at single-cell level. *Nature*, pages 1–9.

Hananeh, A. and Theis, F. (2020). AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution). *bioRxiv*.

Hao, J., Cao, W., Huang, J., Zou, X., and Han, Z. G. (2019). Optimal Gene Filtering for Single-Cell data (OGFSC) - A gene filtering algorithm for single-cell RNA-seq data. *Bioinformatics*, 35(15):2602–2609.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K. J., Rozenblatt-Rosen, O., Dor, Y., Regev, A., and Yanai, I. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(1):77.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3):666–673.

Hayday, A. C. (2019). $\gamma\delta$ T Cell Update: Adaptate Orchestrators of Immune Surveillance. *The Journal of Immunology*, 203(2):311–320.

Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Systems*, 2(4):239–250.

Heinz, D. (2014). Nonparametric mixed membership models. In *Handbook of Mixed Membership Models and Their Applications*, pages 89–116. CRC Press.

Himoudi, N., Morgenstern, D. A., Yan, M., Vernay, B., Saraiva, L., Wu, Y., Cohen, C. J., Gustafsson, K., and Anderson, J. (2012). Human $\gamma\delta$ T Lymphocytes Are Licensed for Professional Antigen Presentation by Interaction with Opsonized Target Cells. *The Journal of Immunology*, 188(4):1708–1716.

Hoffman, P., Satija, R., Papalexi, E., Smibert, P., and Butler, A. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.

Hu, P., Fabyanic, E., Kwon, D. Y., Tang, S., Zhou, Z., and Wu, H. (2017). Dissecting Cell-Type Composition and Activity-Dependent Transcriptional State in Mammalian Brains by Massively Parallel Single-Nucleus RNA-Seq. *Molecular Cell*, 68(5):1006–1015.e7.

Hu, Q. and Greene, C. S. (2019). Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 24:362–373.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.

Huang, Y. (2019). huangyh09/cellSNP: release v0.1.6.

Huang, Y., Heiser, R. A., Detanico, T. O., Getahun, A., Kirchenbaum, G. A., Casper, T. L., Aydintug, M. K., Carding, S. R., Ikuta, K., Huang, H., Cambier, J. C., Wysocki, L. J., O'Brien, R. L., and Born, W. K. (2015). $\gamma\delta$ T cells affect IL-4 production and B-cell tolerance. *Proceedings of the National Academy of Sciences of the United States of America*, 112(1):E39–E48.

Huang, Y., McCarthy, D. J., and Stegle, O. (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biology*, 20(1):273.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hynes, R. O. (2002). Integrins: Bidirectional, allosteric signaling machines. *Cell*, 110(6):673–687.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551.

Jiang, Y., Tang, F., Li, Z., Cui, L., and He, W. (2012). Critical role of $\gamma 4$ chain in the expression of functional V$\gamma$4V$\delta$1 T cell receptor of gastric tumour-infiltrating $\gamma\delta$T lymphocytes. *Scandinavian Journal of Immunology*, 75(1):102–108.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127.

Kabelitz, D., Wesch, D., and He, W. (2007). Perspectives of $\gamma\delta$ T cells in tumor immunology. *Cancer Research*, 67(1):5–8.

Kalinin, A. A., Higgins, G. A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I. D., Najarian, K., and Athey, B. D. (2018). Deep learning in pharmacogenomics: From gene regulation to patient stratification. *Pharmacogenomics*, 19(7):629–650.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., Gate, R. E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L. A., and Ye, C. J. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R. P. (2017). The Drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199.

Karemaker, I. D. and Vermeulen, M. (2018). Single-Cell DNA Methylation Profiling: Technologies and Biological Applications.

Kawasaki, E. S. (2004). Microarrays and the gene expression profile of a single cell. In *Annals of the New York Academy of Sciences*, volume 1020, pages 92–100. New York Academy of Sciences.

Keskar, N. S., Nocedal, J., Tang, P. T. P., Mudigere, D., and Smelyanskiy, M. (2019). On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–2.

Kim, K. T., Lee, H. W., Lee, H. O., Kim, S. C., Seo, Y. J., Chung, W., Eum, H. H., Nam, D. H., Kim, J., Joo, K. M., and Park, W. Y. (2015). Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biology*, 16(1):127.

Kimura, Y., Nagai, N., Tsunekawa, N., Sato-Matsushita, M., Yoshimoto, T., Cua, D. J., Iwakura, Y., Yagita, H., Okada, F., Tahara, H., Saiki, I., Irimura, T., and Hayakawa, Y. (2016). IL-17A-producing CD30+ V$\delta$1 T cells drive inflammation-induced cancer progression. *Cancer Science*, 107(9):1206–1214.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR.

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, page 1.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486.

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). Scmap: Projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15(5):359–362.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

Kobak, D. and Linderman, G. C. (2019). UMAP does not preserve global structure any better than t-SNE when using the same initialization.

Korthauer, K. D., Chu, L. F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222.

Kulig, P., Burkhard, S., Mikita-Geoffroy, J., Croxford, A. L., Hövelmeyer, N., Gyülvészi, G., Gorzelanny, C., Waisman, A., Borsig, L., and Becher, B. (2016). IL17A-mediated endothelial breach promotes metastasis formation. *Cancer Immunology Research*, 4(1):26–32.

Lähnemann, D., Köster, J., Szczurek, E., Mccarthy, D. J., Hicks, S. C., Mark, D., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Emma, M.,

Reinders, M., Ridder, J. D., Saliba, A.-e., and Somarakis, A. (2019). *12 Grand Challenges in Single-Cell Data Science*. Genome Biology.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S. O., Aparicio, S., Baaijens, J., Balvert, M., de Barbanson, B., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T. H., Lelieveldt, B. P., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., de Ridder, J., Saliba, A. E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., and Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):1–35.

Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., Wildberg, A., Gao, D., Fung, H. L., Chen, S., Vijayaraghavan, R., Wong, J., Chen, A., Sheng, X., Kaper, F., Shen, R., Ronaghi, M., Fan, J. B., Wang, W., Chun, J., and Zhang, K. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293):1586–1590.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, 9(1):1–9.

Lieberman, Y., Rokach, L., and Shay, T. (2018). CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS ONE*, 13(10):e0205499.

Lin, P., Troup, M., and Ho, J. W. K. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology*, 18(1):59.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21.

Luo, P., Ding, Y., Lei, X., and Wu, F. X. (2019). DeepDriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Frontiers in Genetics*, 10(JAN):13.

Ma, C., Zhang, Q., Ye, J., Wang, F., Zhang, Y., Wevers, E., Schwartz, T., Hunborg, P., Varvares, M. A., Hoft, D. F., Hsueh, E. C., and Peng, G. (2012). Tumor-Infiltrating $\gamma\delta$ T Lymphocytes Predict Clinical Outcome in Human Breast Cancer. *The Journal of Immunology*, 189(10):5029–5036.

Ma, S., Cheng, Q., Cai, Y., Gong, H., Wu, Y., Yu, X., Shi, L., Wu, D., Dong, C., and Liu, H. (2014). IL-17A produced by $\gamma\delta$ T cells promotes tumor growth in hepatocellular carcinoma. *Cancer Research*, 74(7):1969–1982.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Maniar, A., Zhang, X., Lin, W., Gastman, B. R., Pauza, C. D., Strome, S. E., and Chapoval, A. I. (2010). Human $\gamma\delta$ T lymphocytes induce robust NK cell-mediated antitumor cytotoxicity through CD137 engagement. *Blood*, 116(10):1726–1733.

Mao, C., Mou, X., Zhou, Y., Yuan, G., Xu, C., Liu, H., Zheng, T., Tong, J., Wang, S., and Chen, D. (2014). Tumor-activated TCR$\gamma\delta$+ T cells from gastric cancer patients induce the antitumor immune response of TCR$\alpha\beta$+ T cells via their antigen-presenting cell-like effects. *Journal of Immunology Research*, 2014:593562.

McCarthy, D. J., Rostom, R., Huang, Y., Kunz, D. J., Danecek, P., Bonder, M. J., Hagai, T., Consortium, H., Wang, W., Gaffney, D. J., Simons, B. D., Stegle, O., and Teichmann, S. A. (2018). Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants (under review at Nature Methods). *bioRxiv*, page 413047.

McGinnis, C. S., Murrow, L. M., and Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, 8(4):329–337.e4.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

Meil, M. (2007). Comparing clusterings-an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Mohammadiha, N., Bastiaan Kleijn, W., and Leijon, A. (2013). Gamma hidden Markov model as a probabilistic Nonnegative Matrix Factorization. In *European Signal Processing Conference*.

Montesdeoca, L., Squires, S., and Niranjan, M. (2019). Variational autoencoder for non-negative matrix factorization with exogenous inputs applied to financial data modelling. In *International Symposium on Image and Signal Processing and Analysis, ISPA*, volume 2019-Septe, pages 312–317.

Montoro, D. T., Haber, A. L., Biton, M., Vinarsky, V., Lin, B., Birket, S. E., Yuan, F., Chen, S., Leung, H. M., Villoria, J., Rogel, N., Burgin, G., Tsankov, A. M., Waghray, A., Slyper, M., Waldman, J., Nguyen, L., Dionne, D., Rozenblatt-Rosen, O., Tata, P. R., Mou, H., Shivaraju, M., Bihler, H., Mense, M., Tearney, G. J., Rowe, S. M., Engelhardt, J. F., Regev, A., and Rajagopal, J. (2018). A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature*, 560(7718):319–324.

Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M. A., Carlotti, F., de Koning, E. J., and van Oudenaarden, A. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, 3(4):385–394.e3.

Muto, M., Baghdadi, M., Maekawa, R., Wada, H., and Seino, K. I. (2015). Myeloid molecular characteristics of human $\gamma\delta$ T cells support their acquisition of tumor antigen-presenting capacity. *Cancer Immunology, Immunotherapy*, 64(8):941–949.

Patin, E. C., Soulard, D., Fleury, S., Hassane, M., Dombrowicz, D., Faveeuw, C., Trottein, F., and Paget, C. (2018). Type i IFN receptor signaling controls IL7-dependent accumulation and activity of protumoral IL17A-producing $\gamma\delta$T cells in breast cancer. *Cancer Research*, 78(1):195–204.

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1100.

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181.

Pizzolato, G., Kaminski, H., Tosolini, M., Franchini, D. M., Pont, F., Martins, F., Valle, C., Labourdette, D., Cadot, S., Quillet-Mary, A., Poupot, M., Laurent, C., Ysebaert, L., Meraviglia, S., Dieli, F., Merville, P., Milpied, P., Déchanet-Merville, J., and Fournié, J. J. (2019). Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRV$\delta$1 and TCRV$\delta$2 $\gamma\delta$ T lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 116(24):11906–11915.

Prabhakaran, S., Azizi, E., Carr, A., and Pe'er, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *33rd International Conference on Machine Learning, ICML 2016*, volume 3, pages 1691–1715. International Machine Learning Society (IMLS).

PyTorch (2016). PyTorch.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J., Theis, F. J., Uhlen, M., Van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., and Yosef, N. (2017). The human cell atlas. *eLife*, 6.

Rei, M., Goncalves-Sousa, N., Lanca, T., Thompson, R. G., Mensurado, S., Balkwill, F. R., Kulbe, H., Pennington, D. J., and Silva-Santos, B. (2014). Murine CD27(-) V$\gamma$6(+) $\gamma\delta$ T cells producing IL-17A promote ovarian cancer growth via mobilization of protumor small peritoneal macrophages. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34):E3562–E3570.

Rezende, R. M., Lanser, A. J., Rubino, S., Kuhn, C., Skillin, N., Moreira, T. G., Liu, S., Gabriely, G., David, B. A., Menezes, G. B., and Weiner, H. L. (2018). $\gamma\delta$ T cells control humoral immune response by inducing T follicular helper cell differentiation. *Nature Communications*, 9(1).

Ribot, J. C., DeBarros, A., Pang, D. J., Neves, J. F., Peperzak, V., Roberts, S. J., Girardi, M., Borst, J., Hayday, A. C., Pennington, D. J., and Silva-Santos, B.

(2009). CD27 is a thymic determinant of the balance between interferon-$\gamma$- and interleukin 17-producing $\gamma\delta$ T cell subsets. *Nature Immunology*, 10(4):427–436.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., and Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182.

Ryan, P. L., Sumaria, N., Holland, C. J., Bradford, C. M., Izotova, N., Grandjean, C. L., Jawad, A. S., Bergmeier, L. A., Pennington, D. J., and Born, W. K. (2016). Heterogeneous yet stable V$\delta$2(+) T-cell profiles define distinct cytotoxic effector potentials in healthy human individuals. *Proceedings of the National Academy of Sciences of the United States of America*, 113(50):14378–14383.

Schmidt, M. N., Winther, O., and Kaihansen, L. (2009). Bayesian non-negative matrix factorization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5441, pages 540–547.

Schneider, U., Schwenk, H. ., and Bornkamm, G. (1977). Characterization of EBVgenome negative null and T cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed nonHodgkin lymphoma. *International Journal of Cancer*, 19(5):621–626.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, 8(1):289–317.

Sebestyen, Z., Prinz, I., Déchanet-Merville, J., Silva-Santos, B., and Kuball, J. (2019). Translating gammadelta ($\gamma\delta$) T cells and their receptors into cancer cell therapies. *Nature Reviews Drug Discovery*.

Segerstolpe, A., Palasantza, A., Eliasson, P., Andersson, E.-M., Andreasson, A.-C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M. K., Smith, D. M., Kasper, M., Ammala, C., and Sandberg, R. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, 24(4):593–607.

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., Adiconis, X., Levin, J. Z., Nemesh, J., Goldman, M., McCarroll, S. A., Cepko, C. L., Regev, A., and Sanes, J. R. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166(5):1308–1323.e30.

Silva-Santos, B., Mensurado, S., and Coffelt, S. B. (2019). $\gamma\delta$ T cells: pleiotropic immune effectors with therapeutic potential in cancer. *Nature Reviews Cancer*, 19(7):392–404.

Simões, A. E., Di Lorenzo, B., and Silva-Santos, B. (2018). Molecular determinants of target cell recognition by human $\gamma\delta$ T cells. *Frontiers in Immunology*, 9(APR):929.

Solovjov, D. A., Pluskota, E., and Plow, E. F. (2005). Distinct roles for the $\alpha$ and $\beta$ subunits in the functions of integrin $\alpha$M$\beta$2. *Journal of Biological Chemistry*, 280(2):1336–1345.

Soneson, C. and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261.

Squires, S., Prügel-Bennett, A., and Niranjan, M. (2017). Rank selection in nonnegative matrix factorization using minimum description length. *Neural Computation*, 29(8):2164–2176.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology*, 19(1):224.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21.

Stubbington, M. J., Rozenblatt-Rosen, O., Regev, A., and Teichmann, S. A. (2017). Single-cell transcriptomics to explore the immune system in health and disease. *Science*, 358(6359):58–63.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382.

Tian, L., Dong, X., Freytag, S., Lê Cao, K. A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T. S., Seidi, A., Jabbari, J. S., Naik, S. H., and Ritchie, M. E. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, 16(6):479–487.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., Kazer, S. W., Gaillard, A., Kolb, K. E., Villani, A.-C., Johannessen, C. M., Andreev, A. Y., Van Allen, E. M., Bertagnolli, M., Sorger, P. K., Sullivan, R. J., Flaherty, K. T., Frederick, D. T., Jane-Valbuena, J., Yoon, C. H., Rozenblatt-Rosen, O., Shalek, A. K., Regev, A., and Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196.

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Computational Biology*, 11(6).

Vallejos, C. A., Richardson, S., and Marioni, J. C. (2016). Beyond comparisons of means: Understanding changes in gene expression at the single-cell level. *Genome Biology*, 17(1):70.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nature Methods*, 14(6):565–571.

Van Der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2625.

van Dijk, D., Nainys, J., Sharma, R., Kathail, P., Carr, A. J., Moon, K. R., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. (2017). MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv*.

Virant-Klun, I., Leicht, S., Hughes, C., and Krijgsveld, J. (2016). Identification of maturation-specific proteins by single-cell proteomics of human oocytes. *Molecular and Cellular Proteomics*, 15(8):2616–2627.

Wakita, D., Sumida, K., Iwakura, Y., Nishikawa, H., Ohkuri, T., Chamoto, K., Kitamura, H., and Nishimura, T. (2010). Tumor-infiltrating IL-17-producing $\gamma\delta$ T cells support the progression of tumor by promoting angiogenesis. *European Journal of Immunology*, 40(7):1927–1937.

Wen, L., Pao, W., Wong, F. S., Peng, Q., Craft, J., Zheng, B., Kelsoe, G., Dianda, L., Owen, M. J., and Hayday, A. C. (1996). Germinal center formation, immunoglobulin class switching, and autoantibody production driven by "non $\alpha/\beta$" T cells. *Journal of Experimental Medicine*, 183(5):2271–2282.

Wilson, D. R. and Martinez, T. R. (2003). The general inefficiency of batch training for gradient descent learning. *Neural Networks*, 16(10):1429–1451.

Wolock, S. L., Lopez, R., and Klein, A. M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, 8(4):281–291.e9.

Wu, D., Smyth, G. K., Ritchie, M. E., Law, C. W., Phipson, B., Hu, Y., and Shi, W. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47.

Wu, Y., Kyle-Cezar, F., Woolf, R. T., Naceur-Lombardelli, C., Owen, J., Biswas, D., Lorenc, A., Vantourout, P., Gazinska, P., Grigoriadis, A., Tutt, A., and Hayday, A. (2019). An innate-like V$\delta$1+ $\gamma\delta$ T cell compartment in the human breast is associated with remission in triple-negative breast cancer. *Science Translational Medicine*, 11(513).

Xie, P., Gao, M., Wang, C., Zhang, J., Noel, P., Yang, C., Von Hoff, D., Han, H., Zhang, M. Q., and Lin, W. (2019). SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic acids research*, 47(8):e48.

Xu, C. and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980.

Xu, J., Falconer, C., Nguyen, Q., Crawford, J., McKinnon, B. D., Mortlock, S., Senabouth, A., Andersen, S., Chiu, H. S., Jiang, L., Palpant, N. J., Yang, J., Mueller, M. D., Hewitt, A. W., Pébay, A., Montgomery, G. W., Powell, J. E., and Coin, L. J. (2019). Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biology*, 20(1):290.

Yip, S. H., Sham, P. C., and Wang, J. (2018). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Briefings in Bioinformatics*, 20(4):1583–1589.

Yip, S. H., Wang, P., Kocher, J. P. A., Sham, P. C., and Wang, J. (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic acids research*, 45(22):e179.

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., and Wang, J. (2019a). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, 73(1):130–142.e5.

Zhang, X. F., Ou-Yang, L., Yang, S., Zhao, X. M., Hu, X., Yan, H., and Berger, B. (2019b). EnImpute: Imputing dropout events in single-cell RNA-sequencing data via ensemble learning. *Bioinformatics*, 35(22):4827–4829.

Zhao, Y., Niu, C., and Cui, J. (2018). Gamma-delta ($\gamma\delta$) T Cells: Friend or Foe in Cancer Development. *Journal of Translational Medicine*, 16(1).

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934.

# Appendices

# Appendix A

# Supplementary Tables for Chapter 3

This appendix provides supplementary information of the identified differentially expressed genes and annotated functions for all the data analysed in Chapter 3.

**Table A.1:** List of differentially expressed genes between γδ-T cell subtypes from PBMC.

| gene | HD45 avg_logFC | HD6 avg_logFC | max pval | min pval | cluster |
|---|---|---|---|---|---|
| RPL21 | 3.68e-01 | 4.32e-01 | 2.9e-40 | 2.82e-256 | δ1.1 |
| RPL32 | 3.14e-01 | 5.07e-01 | 2.98e-32 | 5.44e-236 | δ1.1 |
| RPL34 | 2.76e-01 | 4.46e-01 | 4.88e-28 | 1.49e-221 | δ1.1 |
| LEF1 | 9.76e-01 | 1.31e+00 | 8.43e-50 | 1.3e-214 | δ1.1 |
| RPS6 | 2.49e-01 | 4.61e-01 | 1.81e-22 | 7.65e-209 | δ1.1 |
| LDHB | 7.9e-01 | 9.37e-01 | 2.56e-45 | 3.16e-205 | δ1.1 |
| LTB | 6.09e-01 | 1.14e+00 | 2.27e-24 | 1.77e-193 | δ1.1 |
| CCR7 | 5.93e-01 | 1.33e+00 | 9.33e-14 | 1.5e-191 | δ1.1 |
| RPS13 | 3.24e-01 | 5.44e-01 | 2.39e-23 | 1.23e-189 | δ1.1 |
| EEF1A1 | 2.42e-01 | 3.69e-01 | 1.82e-20 | 1.06e-163 | δ1.1 |
| RPS3A | 2.83e-01 | 3.77e-01 | 8.12e-26 | 1.53e-135 | δ1.1 |
| RPL11 | 2.65e-01 | 3.86e-01 | 9.67e-21 | 5.46e-132 | δ1.1 |
| PABPC1 | 3.52e-01 | 6.13e-01 | 8.89e-17 | 2.33e-122 | δ1.1 |
| NELL2 | 6.05e-01 | 7.63e-01 | 1.12e-26 | 2.25e-114 | δ1.1 |
| RTKN2 | 6.47e-01 | 8.88e-01 | 2.42e-18 | 4.34e-106 | δ1.1 |
| ACTN1 | 5.3e-01 | 7.53e-01 | 2.21e-22 | 1.28e-104 | δ1.1 |
| RPS28 | 2.19e-01 | 3.15e-01 | 4.94e-17 | 4.18e-101 | δ1.1 |
| RPS5 | 3.19e-01 | 4.38e-01 | 4.72e-16 | 6.67e-99 | δ1.1 |
| RPL18A | 2.29e-01 | 3.28e-01 | 6.76e-17 | 3.48e-94 | δ1.1 |
| RPL22 | 2.87e-01 | 4.5e-01 | 1.42e-12 | 1.57e-86 | δ1.1 |
| RCAN3 | 4.46e-01 | 7.46e-01 | 1.7e-11 | 1.31e-85 | δ1.1 |
| RP4.594I10.3 | 5.23e-01 | 6.46e-01 | 5.3e-20 | 2.75e-83 | δ1.1 |
| RPL4 | 2.83e-01 | 4.16e-01 | 1.54e-12 | 1.85e-80 | δ1.1 |
| RPS25 | 2.03e-01 | 2.99e-01 | 6.81e-11 | 1.96e-75 | δ1.1 |
| RPLP2 | 2.41e-01 | 2.55e-01 | 2.75e-20 | 3.46e-74 | δ1.1 |
| ID3 | 5.13e-01 | 6.29e-01 | 2.3e-19 | 7.81e-73 | δ1.1 |
| RPS14 | 2.01e-01 | 2.38e-01 | 9.6e-17 | 5.22e-71 | δ1.1 |
| MAL | 5.09e-01 | 5.77e-01 | 1.33e-21 | 5.55e-71 | δ1.1 |
| RPL19 | 2.31e-01 | 2.56e-01 | 1.97e-16 | 4.49e-67 | δ1.1 |
| CAMK4 | 3.79e-01 | 6.3e-01 | 5.15e-07 | 3.31e-66 | δ1.1 |
| TCF7 | 5.35e-01 | 7.14e-01 | 5.02e-12 | 4.86e-64 | δ1.1 |
| RPL30 | 2.65e-01 | 2.8e-01 | 1.39e-15 | 8.42e-63 | δ1.1 |
| PIK3IP1 | 5.81e-01 | 6.84e-01 | 1.57e-14 | 3.03e-61 | δ1.1 |
| LDLRAP1 | 5.34e-01 | 6.65e-01 | 3.19e-12 | 1.05e-60 | δ1.1 |

| | | | | | |
|---|---|---|---|---|---|
| NGFRAP1 | 4.75e-01 | 5.5e-01 | 3.05e-17 | 5.3e-60 | δ1.1 |
| FYB | 5.27e-01 | 6.67e-01 | 5.16e-12 | 4.36e-55 | δ1.1 |
| RPS16 | 2.22e-01 | 2.95e-01 | 7.17e-11 | 1.05e-54 | δ1.1 |
| RPL36 | 2.03e-01 | 2.81e-01 | 5.22e-11 | 1.2e-54 | δ1.1 |
| EEF1B2 | 2.72e-01 | 3.88e-01 | 5.69e-09 | 8.87e-52 | δ1.1 |
| EIF3E | 2.84e-01 | 4.62e-01 | 3.82e-05 | 4.65e-50 | δ1.1 |
| NUCB2 | 6.82e-01 | 6.04e-01 | 1.05e-14 | 4.05e-48 | δ1.1 |
| RPL29 | 2.15e-01 | 2.98e-01 | 2.38e-08 | 2.7e-47 | δ1.1 |
| TRABD2A | 3.15e-01 | 5.11e-01 | 1.89e-06 | 2.3e-46 | δ1.1 |
| RPS9 | 2.13e-01 | 2.54e-01 | 4.29e-11 | 2.4e-46 | δ1.1 |
| SPINT2 | 4.81e-01 | 4.63e-01 | 3.61e-14 | 1.34e-43 | δ1.1 |
| GCSAM | 3.3e-01 | 3.82e-01 | 5.59e-09 | 1.74e-42 | δ1.1 |
| RPL8 | 2.05e-01 | 2.68e-01 | 1.44e-08 | 1.08e-41 | δ1.1 |
| SERINC5 | 2.8e-01 | 4.12e-01 | 3.81e-10 | 1.11e-41 | δ1.1 |
| CHMP7 | 4.25e-01 | 5.33e-01 | 9.81e-10 | 3.52e-39 | δ1.1 |
| HSPB1 | 5.54e-01 | 5.36e-01 | 1.59e-11 | 1.53e-38 | δ1.1 |
| MYC | 3.19e-01 | 5.57e-01 | 7.32e-04 | 3.39e-38 | δ1.1 |
| CCDC109B | 3.41e-01 | 5.09e-01 | 1.04e-06 | 1.99e-37 | δ1.1 |
| TXK | 3.4e-01 | 4.99e-01 | 2.23e-08 | 1.23e-34 | δ1.1 |
| EEF2 | 2.69e-01 | 2.94e-01 | 2.26e-09 | 5.76e-34 | δ1.1 |
| RGCC | 4.45e-01 | 4.94e-01 | 4.47e-10 | 1.06e-29 | δ1.1 |
| CD27 | 4.84e-01 | 4.19e-01 | 3.25e-12 | 1.73e-29 | δ1.1 |
| ARID5B | 5.21e-01 | 4.98e-01 | 1.8e-09 | 1.27e-27 | δ1.1 |
| TMEM123 | 4.57e-01 | 3.7e-01 | 3.94e-11 | 7.89e-26 | δ1.1 |
| PRKCQ.AS1 | 2.71e-01 | 4.09e-01 | 2.27e-04 | 3.86e-25 | δ1.1 |
| BIRC3 | 3.67e-01 | 4.35e-01 | 1.04e-06 | 1.36e-24 | δ1.1 |
| EIF3H | 2.05e-01 | 3.66e-01 | 9.78e-04 | 1.57e-24 | δ1.1 |
| DGKA | 3.28e-01 | 3.94e-01 | 7.66e-06 | 6.52e-24 | δ1.1 |
| NBEAL1 | 2.44e-01 | 3.56e-01 | 1.15e-05 | 8.87e-24 | δ1.1 |
| SUSD3 | 3.47e-01 | 3.74e-01 | 5.62e-07 | 2.61e-23 | δ1.1 |
| MAP3K1 | 3.51e-01 | 3.17e-01 | 1.71e-10 | 5.08e-23 | δ1.1 |
| COTL1 | 4.87e-01 | 4.01e-01 | 3.21e-08 | 4.6e-22 | δ1.1 |
| TMEM66 | 2.21e-01 | 2.97e-01 | 2.62e-04 | 4.98e-22 | δ1.1 |
| FAM102A | 2.29e-01 | 3.23e-01 | 3.18e-05 | 9.66e-22 | δ1.1 |
| FBLN5 | 2.78e-01 | 3.64e-01 | 1.91e-05 | 2.19e-21 | δ1.1 |
| SNHG8 | 3.38e-01 | 3.52e-01 | 4.7e-06 | 3.6e-20 | δ1.1 |
| ISG20 | 2.52e-01 | 3.31e-01 | 1.39e-04 | 3.85e-20 | δ1.1 |
| PFDN5 | 2.14e-01 | 2.44e-01 | 1.25e-06 | 5.04e-20 | δ1.1 |

| | | | | | |
|---|---|---|---|---|---|
| COMMD6 | 2.1e-01 | 2.86e-01 | 3.23e-04 | 1.26e-19 | δ1.1 |
| BTG1 | 2.39e-01 | 2.16e-01 | 1.36e-04 | 1.65e-15 | δ1.1 |
| ANP32B | 2.08e-01 | 3.27e-01 | 5.69e-03 | 3.41e-15 | δ1.1 |
| C1ORF162 | 2.36e-01 | 2.66e-01 | 9.72e-03 | 3.82e-15 | δ1.1 |
| PRKCA | 2.7e-01 | 2.08e-01 | 2.62e-09 | 5.17e-15 | δ1.1 |
| SYPL1 | 2.25e-01 | 2.9e-01 | 3.22e-04 | 6.31e-15 | δ1.1 |
| ST13 | 2.89e-01 | 3.18e-01 | 4.64e-05 | 8.18e-15 | δ1.1 |
| PIM2 | 2.17e-01 | 3.51e-01 | 1.99e-03 | 8.44e-15 | δ1.1 |
| ARHGAP15 | 3.18e-01 | 2.99e-01 | 6.01e-06 | 1.05e-13 | δ1.1 |
| TESPA1 | 2.93e-01 | 2.83e-01 | 4.44e-04 | 1.53e-13 | δ1.1 |
| RIC3 | 2.49e-01 | 2.74e-01 | 5.84e-04 | 2.28e-13 | δ1.1 |
| RPS4Y1 | 2.2e-01 | 2.06e-01 | 7.43e-06 | 1.64e-12 | δ1.1 |
| EIF3L | 2.41e-01 | 2.57e-01 | 3.05e-05 | 1.92e-12 | δ1.1 |
| CCNI | 3.5e-01 | 2.65e-01 | 6.04e-08 | 3.37e-12 | δ1.1 |
| GYPC | 3.04e-01 | 2.72e-01 | 8.56e-05 | 4.38e-12 | δ1.1 |
| AL592284.1 | 3.12e-01 | 2.64e-01 | 1.03e-06 | 4.82e-12 | δ1.1 |
| STMN3 | 3.83e-01 | 2.83e-01 | 1.98e-07 | 1.16e-11 | δ1.1 |
| FOXP1 | 2.15e-01 | 2.68e-01 | 6.43e-03 | 8.75e-11 | δ1.1 |
| FAIM3 | 2.88e-01 | 2.73e-01 | 2.16e-05 | 1.93e-10 | δ1.1 |
| FAM65B | 2.63e-01 | 2.36e-01 | 2.59e-03 | 2.19e-09 | δ1.1 |
| ADSL | 2.57e-01 | 2.27e-01 | 1.12e-04 | 5.65e-09 | δ1.1 |
| ZNF101 | 2.4e-01 | 2.44e-01 | 2.48e-03 | 1.02e-08 | δ1.1 |
| EIF4B | 2.09e-01 | 2.14e-01 | 7.37e-03 | 1.56e-07 | δ1.1 |
| NCF1 | 2.51e-01 | 2.08e-01 | 3.71e-04 | 6.35e-07 | δ1.1 |
| SOX4 | 1.14e+00 | 1.55e+00 | 1.78e-30 | 2.38e-129 | δ1.2 |
| LEF1 | 1.29e+00 | 1.51e+00 | 1.78e-32 | 4.1e-111 | δ1.2 |
| LTB | 1.01e+00 | 1.21e+00 | 8.32e-23 | 9.03e-71 | δ1.2 |
| CHST2 | 5.63e-01 | 8.09e-01 | 3.79e-12 | 1.7e-55 | δ1.2 |
| TCF7 | 7.35e-01 | 9.77e-01 | 7.95e-15 | 2.26e-52 | δ1.2 |
| TMSB10 | 4.13e-01 | 5.49e-01 | 1.22e-12 | 1.47e-51 | δ1.2 |
| CD7 | 7.76e-01 | 8.05e-01 | 6.32e-21 | 5.68e-49 | δ1.2 |
| SMC4 | 7.69e-01 | 1.15e+00 | 1.04e-10 | 5.74e-49 | δ1.2 |
| RTKN2 | 9.04e-01 | 1.02e+00 | 2.68e-13 | 1.51e-45 | δ1.2 |
| TMSB4X | 2.72e-01 | 2.78e-01 | 2.47e-14 | 8.55e-41 | δ1.2 |
| ID3 | 9.03e-01 | 9.51e-01 | 4.11e-16 | 4.9e-38 | δ1.2 |
| SLC25A5 | 6.02e-01 | 7.58e-01 | 4.69e-08 | 1.02e-33 | δ1.2 |
| NUCB2 | 9.98e-01 | 7.96e-01 | 6.53e-20 | 6.73e-33 | δ1.2 |
| HSPB1 | 5.16e-01 | 7.63e-01 | 4.15e-06 | 6.56e-32 | δ1.2 |

| | | | | | |
|---|---|---|---|---|---|
| CHMP7 | 2.78e-01 | 7.92e-01 | 5.5e-03 | 8.12e-31 | δ1.2 |
| IKZF2 | 4.96e-01 | 6.03e-01 | 1.79e-06 | 1.76e-28 | δ1.2 |
| ACTN1 | 7.26e-01 | 7.12e-01 | 1e-14 | 2.72e-28 | δ1.2 |
| SPINT2 | 4.51e-01 | 5.39e-01 | 3.3e-08 | 1.95e-27 | δ1.2 |
| SLFN5 | 4.21e-01 | 6.9e-01 | 8.78e-06 | 2.26e-27 | δ1.2 |
| FYB | 5.07e-01 | 6.85e-01 | 2.34e-06 | 1.67e-26 | δ1.2 |
| AIF1 | 6.31e-01 | 6.99e-01 | 3.45e-15 | 2.6e-25 | δ1.2 |
| RHOH | 7e-01 | 6.89e-01 | 1.69e-10 | 5.57e-25 | δ1.2 |
| PABPC1 | 4.39e-01 | 4.46e-01 | 5.05e-09 | 1.8e-24 | δ1.2 |
| ZNF683 | 1.34e+00 | 8.97e-01 | 4.01e-18 | 4.69e-24 | δ1.2 |
| CHI3L2 | 3.86e-01 | 7.13e-01 | 1.9e-05 | 6.79e-24 | δ1.2 |
| EVL | 5.44e-01 | 5.09e-01 | 3.11e-10 | 1.49e-23 | δ1.2 |
| GCSAM | 3.32e-01 | 5.55e-01 | 5.3e-05 | 2.31e-23 | δ1.2 |
| MARCKSL1 | 6.82e-01 | 6.95e-01 | 1.82e-09 | 1.44e-21 | δ1.2 |
| HIST1H2AC | 4.87e-01 | 6.49e-01 | 1.33e-05 | 2.4e-21 | δ1.2 |
| YBX1 | 2.8e-01 | 4.18e-01 | 6e-05 | 3.4e-20 | δ1.2 |
| H3F3A | 2.1e-01 | 4.17e-01 | 1.1e-02 | 4.29e-20 | δ1.2 |
| RBL2 | 4.38e-01 | 6.04e-01 | 1.87e-04 | 1.03e-19 | δ1.2 |
| ADSL | 3.89e-01 | 5.91e-01 | 1.44e-06 | 4.96e-19 | δ1.2 |
| TOX | 3.24e-01 | 4.69e-01 | 5.58e-05 | 7.79e-19 | δ1.2 |
| MYOM2 | 3.3e-01 | 4.06e-01 | 2.74e-01 | 1.43e-18 | δ1.2 |
| BAZ2B | 4.15e-01 | 4.5e-01 | 1.04e-03 | 7.91e-18 | δ1.2 |
| STMN1 | 7.46e-01 | 5.48e-01 | 4.09e-09 | 4.52e-17 | δ1.2 |
| TMEM123 | 3.92e-01 | 5.43e-01 | 5.72e-04 | 6.82e-17 | δ1.2 |
| CD27 | 6.03e-01 | 5.46e-01 | 8.91e-09 | 1.91e-16 | δ1.2 |
| RPIA | 5e-01 | 6.05e-01 | 1.06e-06 | 4.23e-16 | δ1.2 |
| RCAN3 | 2.24e-01 | 5.48e-01 | 1.94e-04 | 9.44e-16 | δ1.2 |
| EEF2 | 2.54e-01 | 3.27e-01 | 1.96e-03 | 4.77e-15 | δ1.2 |
| FAM65B | 3.24e-01 | 5.07e-01 | 8.77e-03 | 7.88e-15 | δ1.2 |
| CAMK4 | 4.73e-01 | 5.25e-01 | 8.69e-09 | 1.19e-14 | δ1.2 |
| COX6C | 2.86e-01 | 4.11e-01 | 9.82e-03 | 1.81e-14 | δ1.2 |
| C4ORF48 | 2.2e-01 | 5.31e-01 | 1.25e-01 | 1.87e-14 | δ1.2 |
| NGFRAP1 | 3.91e-01 | 4.84e-01 | 1.3e-04 | 2.27e-14 | δ1.2 |
| GYPC | 4.65e-01 | 4.56e-01 | 4.31e-06 | 5.4e-14 | δ1.2 |
| TESPA1 | 3.84e-01 | 4.84e-01 | 9.75e-08 | 9.24e-14 | δ1.2 |
| ITGB2.AS1 | 4.87e-01 | 4.68e-01 | 7.58e-06 | 1.41e-13 | δ1.2 |
| CD79A | 3.01e-01 | 3.88e-01 | 2.33e-08 | 1.46e-13 | δ1.2 |
| 40057 | 3.26e-01 | 4.69e-01 | 9.41e-03 | 1.94e-13 | δ1.2 |

| | | | | | |
|---|---|---|---|---|---|
| ITM2C | 6.05e-01 | 4.26e-01 | 1.59e-07 | 5.57e-13 | δ1.2 |
| CRTAM | 3.37e-01 | 4.75e-01 | 3.15e-03 | 5.63e-13 | δ1.2 |
| HMGA1 | 5.85e-01 | 4.68e-01 | 2.28e-06 | 9.72e-13 | δ1.2 |
| PPP2R5C | 4.91e-01 | 4.34e-01 | 9.36e-06 | 1.31e-12 | δ1.2 |
| MAL | 3.04e-01 | 4.2e-01 | 2.84e-03 | 1.45e-12 | δ1.2 |
| PIM2 | 3.93e-01 | 5.16e-01 | 1.06e-02 | 3.23e-12 | δ1.2 |
| LDHB | 4.67e-01 | 3.77e-01 | 8.54e-08 | 4.15e-12 | δ1.2 |
| STMN3 | 2.19e-01 | 4.67e-01 | 4.92e-02 | 5.4e-12 | δ1.2 |
| STK17B | 5.62e-01 | 4.78e-01 | 5.75e-08 | 6.62e-12 | δ1.2 |
| CCR7 | 3.75e-01 | 4.49e-01 | 6.75e-03 | 4.52e-11 | δ1.2 |
| HMGN1 | 4.71e-01 | 4.05e-01 | 1.17e-07 | 5.97e-11 | δ1.2 |
| MFGE8 | 4.01e-01 | 2.67e-01 | 1.13e-05 | 5.97e-11 | δ1.2 |
| HMGN2 | 2.78e-01 | 3.26e-01 | 6.29e-02 | 6.37e-11 | δ1.2 |
| COTL1 | 5.54e-01 | 4.92e-01 | 8.8e-07 | 7.72e-11 | δ1.2 |
| GNB2L1 | 2.78e-01 | 2.59e-01 | 3.88e-05 | 9.62e-11 | δ1.2 |
| IFITM1 | 4.69e-01 | 4.28e-01 | 1.27e-07 | 3.45e-10 | δ1.2 |
| FKBP1A | 3.52e-01 | 4.02e-01 | 5.96e-05 | 8.22e-10 | δ1.2 |
| INF2 | 3.04e-01 | 2.67e-01 | 4.67e-06 | 9.93e-10 | δ1.2 |
| PRKCH | 5.03e-01 | 4.18e-01 | 3.37e-07 | 9.95e-10 | δ1.2 |
| CCNI | 3.63e-01 | 3.63e-01 | 4.64e-04 | 1.32e-09 | δ1.2 |
| EPHX2 | 3.06e-01 | 3.16e-01 | 3.26e-04 | 1.58e-09 | δ1.2 |
| ILF3.AS1 | 2.11e-01 | 4.55e-01 | 4.19e-02 | 2.77e-09 | δ1.2 |
| FAM173A | 2.42e-01 | 4.46e-01 | 2.58e-03 | 3.02e-09 | δ1.2 |
| LIMD2 | 2.12e-01 | 3.4e-01 | 9.24e-02 | 3.78e-09 | δ1.2 |
| ARID5B | 5.75e-01 | 4.83e-01 | 3.66e-08 | 4.74e-09 | δ1.2 |
| CLEC11A | 3.11e-01 | 2.54e-01 | 2.85e-09 | 5.44e-09 | δ1.2 |
| PARP1 | 3.06e-01 | 4.13e-01 | 1.89e-03 | 8.04e-09 | δ1.2 |
| RASGRP2 | 2.73e-01 | 4.37e-01 | 3.92e-02 | 8.99e-09 | δ1.2 |
| FCGRT | 4.03e-01 | 3.71e-01 | 1.87e-05 | 1.05e-08 | δ1.2 |
| FAM102A | 2.73e-01 | 3.4e-01 | 2.19e-03 | 1.15e-08 | δ1.2 |
| LDLRAP1 | 5.7e-01 | 3.83e-01 | 6.77e-06 | 1.42e-08 | δ1.2 |
| NSMCE1 | 3.52e-01 | 3.44e-01 | 1.84e-03 | 1.52e-08 | δ1.2 |
| ACAP1 | 2.14e-01 | 3.49e-01 | 1.83e-01 | 2.11e-08 | δ1.2 |
| DGKA | 3.47e-01 | 3.77e-01 | 5.08e-03 | 2.52e-08 | δ1.2 |
| IRF2BP2 | 2.22e-01 | 2.95e-01 | 3.35e-02 | 2.66e-08 | δ1.2 |
| TCF12 | 2.44e-01 | 3.24e-01 | 2.15e-03 | 2.82e-08 | δ1.2 |
| ELF1 | 3.92e-01 | 4.28e-01 | 1.28e-02 | 4.03e-08 | δ1.2 |
| RP11.347P5.1 | 5.23e-01 | 4.31e-01 | 9.1e-05 | 4.35e-08 | δ1.2 |

| | | | | | |
|---|---|---|---|---|---|
| HIST1H2BK | 3.06e-01 | 3.37e-01 | 5.17e-04 | 5.39e-08 | δ1.2 |
| TNRC6B | 2.87e-01 | 3.7e-01 | 7.83e-02 | 7.24e-08 | δ1.2 |
| EIF5A | 2.17e-01 | 3.11e-01 | 5.3e-02 | 9.05e-08 | δ1.2 |
| SERINC5 | 2.67e-01 | 2.22e-01 | 7.39e-03 | 1.02e-07 | δ1.2 |
| CCDC69 | 2.6e-01 | 3.31e-01 | 1.78e-02 | 1.06e-07 | δ1.2 |
| CD52 | 3.96e-01 | 2.01e-01 | 7.78e-06 | 1.11e-07 | δ1.2 |
| NOSIP | 4.48e-01 | 3.87e-01 | 9.51e-06 | 2.25e-07 | δ1.2 |
| SUSD3 | 4.87e-01 | 2.84e-01 | 3.35e-05 | 4.49e-07 | δ1.2 |
| CLDND1 | 2.88e-01 | 3.53e-01 | 3.34e-02 | 6.54e-07 | δ1.2 |
| 38961 | 3.76e-01 | 3.27e-01 | 2.91e-04 | 7.25e-07 | δ1.2 |
| MLLT11 | 3.5e-01 | 2.25e-01 | 9.17e-06 | 7.44e-07 | δ1.2 |
| EIF3L | 4.09e-01 | 2.76e-01 | 2.53e-05 | 1e-06 | δ1.2 |
| CXXC5 | 4.69e-01 | 2.42e-01 | 7.78e-05 | 1.26e-06 | δ1.2 |
| LPGAT1 | 2.01e-01 | 2.34e-01 | 3.86e-02 | 1.36e-06 | δ1.2 |
| ATXN10 | 2.06e-01 | 2.08e-01 | 4.11e-03 | 1.38e-06 | δ1.2 |
| PRMT1 | 5.07e-01 | 3.15e-01 | 2.38e-05 | 1.39e-06 | δ1.2 |
| BIRC2 | 2.5e-01 | 3.25e-01 | 4.16e-02 | 1.8e-06 | δ1.2 |
| CCDC109B | 3.24e-01 | 3.65e-01 | 4.83e-03 | 2.13e-06 | δ1.2 |
| TUBA1A | 2.36e-01 | 3.46e-01 | 1.09e-01 | 2.17e-06 | δ1.2 |
| CTD.2020K17.1 | 4.5e-01 | 2.81e-01 | 2.71e-05 | 2.52e-06 | δ1.2 |
| FNBP1 | 2.06e-01 | 3.51e-01 | 8.14e-02 | 2.56e-06 | δ1.2 |
| ETFB | 3.92e-01 | 3.61e-01 | 6.16e-05 | 2.64e-06 | δ1.2 |
| CCDC57 | 3.67e-01 | 3.46e-01 | 6.35e-05 | 2.65e-06 | δ1.2 |
| NIN | 4.82e-01 | 3.46e-01 | 8.08e-06 | 2.91e-06 | δ1.2 |
| H2AFY | 4.59e-01 | 3.62e-01 | 3.97e-05 | 3.14e-06 | δ1.2 |
| C16ORF54 | 3.33e-01 | 4.01e-01 | 3.41e-03 | 3.21e-06 | δ1.2 |
| 37135 | 2.83e-01 | 2.75e-01 | 4.24e-02 | 6.61e-06 | δ1.2 |
| ALKBH7 | 2.2e-01 | 2.37e-01 | 1.46e-02 | 7.75e-06 | δ1.2 |
| BEX4 | 4.74e-01 | 2.49e-01 | 1.36e-04 | 9.86e-06 | δ1.2 |
| AP1S2 | 2.36e-01 | 3.02e-01 | 8.07e-02 | 1.16e-05 | δ1.2 |
| ZSCAN18 | 3.19e-01 | 2.04e-01 | 2.36e-04 | 1.16e-05 | δ1.2 |
| RTN4 | 2.39e-01 | 2.91e-01 | 1.76e-03 | 1.26e-05 | δ1.2 |
| CALM2 | 2.39e-01 | 2.53e-01 | 3.71e-03 | 1.38e-05 | δ1.2 |
| CCNG2 | 2.43e-01 | 2.84e-01 | 1.7e-02 | 1.53e-05 | δ1.2 |
| UBL7 | 2.02e-01 | 2.04e-01 | 1.01e-02 | 2.01e-05 | δ1.2 |
| MDM4 | 5.83e-01 | 3.25e-01 | 2.71e-05 | 2.2e-05 | δ1.2 |
| ADD3 | 2.73e-01 | 3.4e-01 | 3.59e-03 | 2.48e-05 | δ1.2 |
| CXCR3 | 4.69e-01 | 3.1e-01 | 6.69e-04 | 2.81e-05 | δ1.2 |

| | | | | | |
|---|---|---|---|---|---|
| ETS1 | 2.41e-01 | 2.9e-01 | 6.91e-02 | 2.88e-05 | δ1.2 |
| CCDC167 | 2.18e-01 | 3.12e-01 | 4.63e-02 | 3.08e-05 | δ1.2 |
| N4BP2 | 3.26e-01 | 2.44e-01 | 1.23e-04 | 3.26e-05 | δ1.2 |
| RAB37 | 3.76e-01 | 3.25e-01 | 4.36e-03 | 5.25e-05 | δ1.2 |
| SDCCAG8 | 4.04e-01 | 2.35e-01 | 3.62e-04 | 5.37e-05 | δ1.2 |
| LAT | 3.59e-01 | 2.74e-01 | 5.7e-04 | 6.08e-05 | δ1.2 |
| ATP5C1 | 3.01e-01 | 2.85e-01 | 4.18e-05 | 6.22e-05 | δ1.2 |
| ESYT1 | 2.91e-01 | 2.96e-01 | 1.27e-04 | 6.34e-05 | δ1.2 |
| C6ORF48 | 3.82e-01 | 2.6e-01 | 6.11e-04 | 6.63e-05 | δ1.2 |
| PRPSAP2 | 3.53e-01 | 2.39e-01 | 2.58e-03 | 7.03e-05 | δ1.2 |
| ZFX | 2.45e-01 | 2.75e-01 | 7.32e-02 | 7.69e-05 | δ1.2 |
| ANP32A | 3.02e-01 | 2.19e-01 | 4.67e-03 | 8.05e-05 | δ1.2 |
| CD84 | 3.71e-01 | 2.83e-01 | 1.54e-04 | 9.48e-05 | δ1.2 |
| SMARCA4 | 3.21e-01 | 2.46e-01 | 8.77e-04 | 1.04e-04 | δ1.2 |
| PGLS | 2.21e-01 | 2.91e-01 | 1.08e-01 | 1.25e-04 | δ1.2 |
| NRROS | 2.74e-01 | 2.56e-01 | 3.27e-04 | 1.32e-04 | δ1.2 |
| STT3B | 3.61e-01 | 2.65e-01 | 7.16e-04 | 1.34e-04 | δ1.2 |
| IP6K2 | 3.82e-01 | 3.14e-01 | 1.28e-03 | 1.38e-04 | δ1.2 |
| QARS | 3.63e-01 | 2.72e-01 | 4.96e-04 | 1.45e-04 | δ1.2 |
| ARGLU1 | 3.42e-01 | 2.34e-01 | 3.2e-03 | 1.49e-04 | δ1.2 |
| ARHGEF1 | 2.8e-01 | 2.72e-01 | 2.52e-03 | 1.71e-04 | δ1.2 |
| NAP1L4 | 3.17e-01 | 2.9e-01 | 3.26e-04 | 1.89e-04 | δ1.2 |
| TLE4 | 2.43e-01 | 3.08e-01 | 3.95e-03 | 1.99e-04 | δ1.2 |
| DCAF7 | 2.65e-01 | 2.13e-01 | 5.51e-02 | 2.01e-04 | δ1.2 |
| MLXIP | 2.08e-01 | 2.39e-01 | 9.19e-03 | 2.01e-04 | δ1.2 |
| B3GNT2 | 3.57e-01 | 2.92e-01 | 8.53e-04 | 2.02e-04 | δ1.2 |
| PPP1CC | 2.69e-01 | 2.38e-01 | 3.6e-02 | 2.14e-04 | δ1.2 |
| LSM2 | 2.34e-01 | 2.22e-01 | 3.27e-02 | 2.18e-04 | δ1.2 |
| PTEN | 3.21e-01 | 2.58e-01 | 1.08e-02 | 2.43e-04 | δ1.2 |
| TRAF3IP3 | 3.38e-01 | 2.77e-01 | 3.98e-04 | 2.5e-04 | δ1.2 |
| ZNF428 | 2.43e-01 | 2.13e-01 | 8.28e-02 | 2.62e-04 | δ1.2 |
| CNN2 | 2.3e-01 | 3e-01 | 1.39e-01 | 2.94e-04 | δ1.2 |
| TIMM10 | 2.78e-01 | 2.04e-01 | 2.98e-02 | 3.01e-04 | δ1.2 |
| BIRC3 | 2.08e-01 | 2.76e-01 | 2.51e-02 | 3.44e-04 | δ1.2 |
| TXK | 4.04e-01 | 2.52e-01 | 1.13e-03 | 3.49e-04 | δ1.2 |
| GTF3A | 2.14e-01 | 2.21e-01 | 1.82e-02 | 3.58e-04 | δ1.2 |
| R3HDM4 | 2.21e-01 | 2.57e-01 | 8.16e-02 | 3.61e-04 | δ1.2 |
| CSK | 2.96e-01 | 2.39e-01 | 5.08e-03 | 3.74e-04 | δ1.2 |

| | | | | | |
|---|---|---|---|---|---|
| DCK | 2.48e-01 | 2.79e-01 | 1.52e-03 | 5.54e-04 | δ1.2 |
| CCM2 | 4.28e-01 | 2.46e-01 | 3.84e-04 | 6.18e-04 | δ1.2 |
| UCP2 | 2.16e-01 | 2.62e-01 | 8.72e-02 | 7.17e-04 | δ1.2 |
| PPP4R2 | 3.78e-01 | 2.53e-01 | 7.97e-04 | 7.73e-04 | δ1.2 |
| HDAC2 | 2.7e-01 | 2.1e-01 | 5.92e-02 | 1.18e-03 | δ1.2 |
| HDAC7 | 2.95e-01 | 2.57e-01 | 5.63e-03 | 1.32e-03 | δ1.2 |
| CCT8 | 2.07e-01 | 2.12e-01 | 1.17e-01 | 1.44e-03 | δ1.2 |
| UBE2V2 | 3.61e-01 | 2.11e-01 | 1.82e-03 | 1.53e-03 | δ1.2 |
| EIF4G2 | 2.33e-01 | 2.03e-01 | 4.32e-02 | 1.63e-03 | δ1.2 |
| STK17A | 3.79e-01 | 2.54e-01 | 2.94e-03 | 1.71e-03 | δ1.2 |
| EPB41 | 2.04e-01 | 2.4e-01 | 1.66e-01 | 1.99e-03 | δ1.2 |
| SETD5.AS1 | 2.7e-01 | 2.12e-01 | 6.07e-03 | 2.1e-03 | δ1.2 |
| TRAF5 | 2.88e-01 | 2.26e-01 | 3.48e-03 | 2.44e-03 | δ1.2 |
| CCSER2 | 2.35e-01 | 2e-01 | 2.55e-03 | 2.83e-03 | δ1.2 |
| IFNAR2 | 2.01e-01 | 2.01e-01 | 6.33e-03 | 5.37e-03 | δ1.2 |
| LUC7L3 | 3e-01 | 2.33e-01 | 3.86e-02 | 5.45e-03 | δ1.2 |
| USP11 | 2.01e-01 | 2.2e-01 | 7.75e-03 | 5.69e-03 | δ1.2 |
| TPR | 2.79e-01 | 2.85e-01 | 1.13e-02 | 6.44e-03 | δ1.2 |
| AKIRIN1 | 2.81e-01 | 2.1e-01 | 3.44e-03 | 6.6e-03 | δ1.2 |
| FBXL3 | 2.12e-01 | 2.1e-01 | 1.33e-02 | 7.73e-03 | δ1.2 |
| PPM1M | 2.62e-01 | 2.11e-01 | 3.11e-02 | 9.66e-03 | δ1.2 |
| SNORA76 | 2.13e-01 | 2.26e-01 | 2.96e-01 | 1.85e-02 | δ1.2 |
| GRK6 | 2.1e-01 | 2.27e-01 | 1.04e-01 | 1.93e-02 | δ1.2 |
| CHD4 | 2.06e-01 | 2.22e-01 | 1.69e-01 | 3.48e-02 | δ1.2 |
| GNLY | 1.97e+00 | 1.44e+00 | 7.2e-233 | 0e+00 | δ2.1 |
| FGFBP2 | 2.04e+00 | 1.56e+00 | 1.19e-210 | 0e+00 | δ2.1 |
| GZMH | 1.84e+00 | 1.36e+00 | 7.01e-197 | 0e+00 | δ2.1 |
| NKG7 | 9.68e-01 | 8.62e-01 | 2.64e-183 | 0e+00 | δ2.1 |
| B2M | 2.35e-01 | 2.57e-01 | 2.6e-90 | 9.85e-299 | δ2.1 |
| GZMB | 1.9e+00 | 1.24e+00 | 2.21e-209 | 1.98e-245 | δ2.1 |
| PRF1 | 9.49e-01 | 7.88e-01 | 2.13e-117 | 3.23e-200 | δ2.1 |
| CST7 | 6.27e-01 | 5.99e-01 | 1.24e-70 | 6.71e-190 | δ2.1 |
| KLRD1 | 6.04e-01 | 7.07e-01 | 3.87e-49 | 9.89e-165 | δ2.1 |
| CCL4 | 9.78e-01 | 7.3e-01 | 3.32e-81 | 3.62e-156 | δ2.1 |
| CCL5 | 3.42e-01 | 3.66e-01 | 1.04e-41 | 9.48e-153 | δ2.1 |
| CX3CR1 | 9.02e-01 | 8.06e-01 | 3.72e-77 | 4.3e-145 | δ2.1 |
| FCGR3A | 1.39e+00 | 7.48e-01 | 4.11e-103 | 7.42e-133 | δ2.1 |
| HLA.C | 2.83e-01 | 3.14e-01 | 1.28e-41 | 4.27e-126 | δ2.1 |

| | | | | | |
|---|---|---|---|---|---|
| GPR56 | 9.16e-01 | 6.93e-01 | 4.11e-77 | 2.98e-112 | δ2.1 |
| S100A4 | 3.11e-01 | 3.94e-01 | 9.09e-31 | 2.61e-108 | δ2.1 |
| HLA.B | 2.36e-01 | 2.23e-01 | 2.93e-44 | 9.91e-94 | δ2.1 |
| SH3BGRL3 | 4.34e-01 | 3.91e-01 | 3.35e-49 | 3.68e-89 | δ2.1 |
| TTC38 | 6.58e-01 | 5.55e-01 | 2.33e-47 | 4.64e-88 | δ2.1 |
| PRSS23 | 7.76e-01 | 5.49e-01 | 5.91e-70 | 1.3e-84 | δ2.1 |
| GZMA | 2.94e-01 | 3.29e-01 | 9.73e-24 | 9.93e-83 | δ2.1 |
| SRGN | 4.09e-01 | 4.07e-01 | 9.77e-33 | 6.75e-82 | δ2.1 |
| SPON2 | 7.73e-01 | 6.09e-01 | 5.99e-46 | 1.25e-78 | δ2.1 |
| S1PR5 | 6.75e-01 | 5.64e-01 | 2.47e-45 | 1.95e-75 | δ2.1 |
| TBX21 | 6.76e-01 | 5.62e-01 | 2.7e-41 | 3.55e-75 | δ2.1 |
| ZEB2 | 6.49e-01 | 5.09e-01 | 4.07e-48 | 1.08e-68 | δ2.1 |
| ITGB2 | 4.65e-01 | 3.87e-01 | 4.26e-34 | 3.1e-68 | δ2.1 |
| PLEK | 6.63e-01 | 5.19e-01 | 1.48e-47 | 3.13e-66 | δ2.1 |
| CD247 | 4.47e-01 | 4.34e-01 | 1.93e-23 | 2.72e-63 | δ2.1 |
| HLA.DPB1 | 4.05e-01 | 4.72e-01 | 5.63e-22 | 1.09e-62 | δ2.1 |
| APOBEC3G | 3.61e-01 | 4.87e-01 | 1.86e-14 | 1.44e-62 | δ2.1 |
| ARPC2 | 4.31e-01 | 3.21e-01 | 1.59e-36 | 3.64e-62 | δ2.1 |
| LGALS1 | 7.35e-01 | 6.62e-01 | 5.97e-28 | 3.59e-61 | δ2.1 |
| EFHD2 | 5.46e-01 | 5.05e-01 | 3.07e-29 | 3.69e-60 | δ2.1 |
| PLAC8 | 4.73e-01 | 4.17e-01 | 2.18e-27 | 6.65e-59 | δ2.1 |
| ITGB1 | 2.35e-01 | 5.3e-01 | 8.93e-10 | 1.24e-57 | δ2.1 |
| ASCL2 | 4.85e-01 | 4.16e-01 | 1.07e-24 | 1.51e-56 | δ2.1 |
| CYBA | 3.74e-01 | 2.97e-01 | 2.41e-35 | 2.21e-56 | δ2.1 |
| HOPX | 5.88e-01 | 4.32e-01 | 2.07e-31 | 1.56e-53 | δ2.1 |
| IFITM2 | 3.71e-01 | 3.49e-01 | 1.98e-21 | 1.22e-51 | δ2.1 |
| PYHIN1 | 2.72e-01 | 4.6e-01 | 3.05e-07 | 1.43e-51 | δ2.1 |
| ZNF683 | 2.59e-01 | 5.42e-01 | 2.71e-09 | 5.45e-51 | δ2.1 |
| FGR | 5.35e-01 | 4.53e-01 | 7.48e-27 | 2.04e-50 | δ2.1 |
| LITAF | 5.46e-01 | 3.42e-01 | 3.37e-40 | 5.59e-50 | δ2.1 |
| HLA.E | 2.04e-01 | 2.41e-01 | 3.84e-15 | 1.02e-49 | δ2.1 |
| C1ORF21 | 4.94e-01 | 4.58e-01 | 3.15e-23 | 5.65e-49 | δ2.1 |
| CTSW | 3.52e-01 | 3.1e-01 | 7.85e-27 | 1.29e-48 | δ2.1 |
| CD3D | 2.95e-01 | 2.79e-01 | 1.12e-19 | 1.75e-48 | δ2.1 |
| MYO1F | 3.55e-01 | 3.81e-01 | 3.47e-16 | 6.29e-48 | δ2.1 |
| PFN1 | 3.32e-01 | 2.3e-01 | 2.09e-36 | 2.39e-45 | δ2.1 |
| RAP1B | 5.72e-01 | 3.18e-01 | 9.76e-42 | 3.46e-45 | δ2.1 |
| AKR1C3 | 4.15e-01 | 3.94e-01 | 1.92e-21 | 4.44e-44 | δ2.1 |

| | | | | | |
|---|---|---|---|---|---|
| CD99 | 3.62e-01 | 3.04e-01 | 1.03e-21 | 2.04e-43 | δ2.1 |
| CTSC | 4.48e-01 | 3.64e-01 | 7.7e-22 | 7.36e-40 | δ2.1 |
| CLIC1 | 2.54e-01 | 3.12e-01 | 1.2e-10 | 9.64e-40 | δ2.1 |
| ABI3 | 5.47e-01 | 3.99e-01 | 2.64e-25 | 1.4e-39 | δ2.1 |
| ADRB2 | 3.66e-01 | 4.11e-01 | 1.68e-14 | 3.6e-39 | δ2.1 |
| C12ORF75 | 5.75e-01 | 3.64e-01 | 1.22e-33 | 5.13e-37 | δ2.1 |
| FCRL6 | 5.62e-01 | 3.62e-01 | 1.35e-26 | 1.13e-36 | δ2.1 |
| APMAP | 4e-01 | 3.52e-01 | 3.9e-18 | 4.29e-36 | δ2.1 |
| TYROBP | 5.06e-01 | 5.08e-01 | 6.21e-12 | 6.55e-36 | δ2.1 |
| SPN | 3.38e-01 | 3.54e-01 | 2.46e-10 | 2.57e-35 | δ2.1 |
| SAMD3 | 3.14e-01 | 3.3e-01 | 6.59e-12 | 3.58e-35 | δ2.1 |
| CHST12 | 3.41e-01 | 3.44e-01 | 1.11e-13 | 1.65e-34 | δ2.1 |
| THEMIS2 | 2.69e-01 | 3.39e-01 | 6.33e-09 | 2.8e-34 | δ2.1 |
| TNFRSF1B | 4.33e-01 | 3.71e-01 | 3.64e-18 | 8.35e-34 | δ2.1 |
| ACTB | 3.42e-01 | 2.5e-01 | 3.24e-26 | 3.33e-32 | δ2.1 |
| DOK2 | 3.67e-01 | 3.41e-01 | 4.89e-14 | 6.58e-32 | δ2.1 |
| BIN2 | 3.06e-01 | 2.57e-01 | 4.17e-14 | 1.68e-31 | δ2.1 |
| ACTG1 | 2.89e-01 | 2.82e-01 | 4.06e-11 | 3.25e-31 | δ2.1 |
| PATL2 | 4.46e-01 | 3.67e-01 | 3.73e-18 | 9.64e-30 | δ2.1 |
| CAP1 | 2.83e-01 | 2.52e-01 | 1.31e-10 | 2.97e-29 | δ2.1 |
| CCND3 | 3.05e-01 | 2.61e-01 | 2.08e-10 | 3.1e-29 | δ2.1 |
| PPP1CA | 3.21e-01 | 2.44e-01 | 9.16e-14 | 5.69e-29 | δ2.1 |
| CLIC3 | 3.89e-01 | 3.68e-01 | 3.83e-14 | 8.4e-28 | δ2.1 |
| ID2 | 3.97e-01 | 3.13e-01 | 1.12e-15 | 2.54e-27 | δ2.1 |
| MATK | 2.27e-01 | 2.95e-01 | 1.54e-05 | 5.22e-27 | δ2.1 |
| GZMM | 2.93e-01 | 2.69e-01 | 3.16e-11 | 1.22e-26 | δ2.1 |
| LPCAT1 | 2.48e-01 | 2.96e-01 | 1.94e-07 | 1.97e-26 | δ2.1 |
| GNPTAB | 3.09e-01 | 3.05e-01 | 3.7e-09 | 2.26e-26 | δ2.1 |
| GTF3C1 | 2.4e-01 | 3.11e-01 | 3.69e-07 | 1.21e-25 | δ2.1 |
| CD63 | 3.71e-01 | 2.89e-01 | 2e-13 | 3.68e-25 | δ2.1 |
| CD3G | 3.61e-01 | 2.35e-01 | 1.87e-18 | 5.95e-25 | δ2.1 |
| DBI | 2.84e-01 | 2.44e-01 | 2.7e-09 | 1.21e-24 | δ2.1 |
| EMP3 | 2.27e-01 | 2.26e-01 | 1.03e-09 | 1.61e-24 | δ2.1 |
| MIAT | 2.43e-01 | 2.8e-01 | 4.3e-07 | 1.73e-24 | δ2.1 |
| CD300A | 2.65e-01 | 3.18e-01 | 3.11e-06 | 2.28e-24 | δ2.1 |
| UBE2F | 3.61e-01 | 2.86e-01 | 2.31e-16 | 3.05e-24 | δ2.1 |
| FLNA | 2.5e-01 | 2.57e-01 | 2.99e-06 | 2.59e-23 | δ2.1 |
| MT2A | 3.29e-01 | 2.98e-01 | 3.12e-10 | 7.45e-23 | δ2.1 |

| | | | | | |
|---|---|---|---|---|---|
| TPST2 | 3.02e-01 | 2.87e-01 | 6.24e-09 | 9.91e-23 | δ2.1 |
| BSG | 2.2e-01 | 2.03e-01 | 3.37e-06 | 1.08e-22 | δ2.1 |
| SH3BP5 | 2.21e-01 | 2.89e-01 | 2.69e-05 | 2.2e-22 | δ2.1 |
| DYNLL1 | 2.43e-01 | 2.69e-01 | 1.77e-06 | 3.55e-22 | δ2.1 |
| RHOC | 3.57e-01 | 2.98e-01 | 3.02e-13 | 3.85e-22 | δ2.1 |
| FAM49B | 2.45e-01 | 2.38e-01 | 3.45e-07 | 5.26e-22 | δ2.1 |
| CTSD | 3.45e-01 | 2.2e-01 | 3.84e-12 | 6.39e-22 | δ2.1 |
| GNG2 | 3.61e-01 | 2.61e-01 | 2.06e-12 | 6.96e-22 | δ2.1 |
| CD97 | 3.3e-01 | 2.61e-01 | 3.7e-12 | 1.04e-21 | δ2.1 |
| ITGAM | 3.21e-01 | 2.51e-01 | 4.38e-13 | 2.81e-21 | δ2.1 |
| MYO1G | 4.03e-01 | 2.54e-01 | 1.4e-15 | 3.21e-21 | δ2.1 |
| PLEKHG3 | 2.48e-01 | 2.49e-01 | 6.01e-08 | 5.57e-21 | δ2.1 |
| LSP1 | 3.63e-01 | 2.16e-01 | 2.53e-20 | 5.97e-21 | δ2.1 |
| C5ORF56 | 2.88e-01 | 2.7e-01 | 3.65e-07 | 8.85e-21 | δ2.1 |
| ARL6IP5 | 2.91e-01 | 2.07e-01 | 4.41e-09 | 2.28e-20 | δ2.1 |
| GSTP1 | 2.54e-01 | 2.16e-01 | 2.3e-07 | 3.64e-20 | δ2.1 |
| ACTR3 | 2.36e-01 | 2.02e-01 | 2.61e-06 | 3.87e-20 | δ2.1 |
| TOB1 | 3.13e-01 | 2.88e-01 | 3.32e-08 | 3.92e-20 | δ2.1 |
| SLC9A3R1 | 3.85e-01 | 2.31e-01 | 2.41e-17 | 5.19e-20 | δ2.1 |
| PSAP | 3.9e-01 | 2.09e-01 | 1.41e-15 | 1.33e-19 | δ2.1 |
| WDR1 | 2.44e-01 | 2.11e-01 | 4.44e-06 | 1.59e-19 | δ2.1 |
| PTGER2 | 3.59e-01 | 2.73e-01 | 2.38e-11 | 1.62e-19 | δ2.1 |
| CD320 | 3.99e-01 | 2.41e-01 | 5.23e-15 | 4.39e-19 | δ2.1 |
| PTGDR | 2.35e-01 | 2.69e-01 | 3.74e-06 | 7.55e-19 | δ2.1 |
| PPP1R18 | 2.47e-01 | 2.05e-01 | 6.43e-07 | 7.15e-18 | δ2.1 |
| TMEM173 | 2.16e-01 | 2.34e-01 | 2.94e-05 | 1.16e-17 | δ2.1 |
| CRIP1 | 3.86e-01 | 2.57e-01 | 2e-15 | 2.78e-17 | δ2.1 |
| ITGAL | 3.4e-01 | 2.13e-01 | 1.47e-11 | 5.02e-17 | δ2.1 |
| MAPK1 | 2.73e-01 | 2.04e-01 | 1.11e-06 | 8.9e-17 | δ2.1 |
| ABHD17A | 3.69e-01 | 2.34e-01 | 2.97e-13 | 2.66e-16 | δ2.1 |
| AOAH | 3.01e-01 | 2.33e-01 | 2.48e-10 | 3.11e-16 | δ2.1 |
| LAIR1 | 3.68e-01 | 2.24e-01 | 2.21e-13 | 3.36e-16 | δ2.1 |
| AC092580.4 | 2.33e-01 | 2.43e-01 | 8.52e-08 | 3.72e-16 | δ2.1 |
| APOBEC3C | 3.8e-01 | 2.64e-01 | 5.56e-11 | 6.47e-16 | δ2.1 |
| PLEKHF1 | 2.62e-01 | 2.16e-01 | 2.89e-07 | 6.02e-14 | δ2.1 |
| TPM4 | 2.4e-01 | 2.02e-01 | 1.94e-05 | 2.6e-13 | δ2.1 |
| STARD3NL | 2.18e-01 | 2.08e-01 | 1.9e-06 | 5.74e-13 | δ2.1 |
| GZMK | 1.13e+00 | 1.48e+00 | 1.61e-182 | 0e+00 | δ2.2 |

| | | | | | |
|---|---|---|---|---|---|
| CD74 | 4.65e-01 | 7.58e-01 | 8.14e-28 | 1.25e-152 | δ2.2 |
| KLRB1 | 4.78e-01 | 3.91e-01 | 8.69e-32 | 5.25e-92 | δ2.2 |
| XCL1 | 3.04e-01 | 5.28e-01 | 1.33e-09 | 3.58e-63 | δ2.2 |
| CXCR6 | 4.83e-01 | 4.7e-01 | 5.87e-24 | 1.13e-52 | δ2.2 |
| DUSP2 | 4.53e-01 | 3.94e-01 | 2.73e-20 | 6.91e-47 | δ2.2 |
| LYAR | 3.25e-01 | 3.25e-01 | 5.39e-14 | 1.09e-39 | δ2.2 |
| CEBPD | 3.17e-01 | 4.01e-01 | 3.69e-11 | 1.81e-37 | δ2.2 |
| CD160 | 2.55e-01 | 3.69e-01 | 1.06e-06 | 7.67e-34 | δ2.2 |
| DPP4 | 3.42e-01 | 2.77e-01 | 1.82e-14 | 2.05e-31 | δ2.2 |
| CCR5 | 3.11e-01 | 2.86e-01 | 5.87e-13 | 6.96e-29 | δ2.2 |
| KLRG1 | 2.39e-01 | 2.41e-01 | 1.13e-10 | 8.34e-29 | δ2.2 |
| HLA.DRB1 | 2.6e-01 | 3.8e-01 | 8.07e-06 | 1.65e-27 | δ2.2 |
| IL7R | 4.08e-01 | 2.22e-01 | 1.96e-22 | 2.59e-22 | δ2.2 |
| HLA.DMA | 2.2e-01 | 2.36e-01 | 5.27e-07 | 6.66e-22 | δ2.2 |
| IFNGR1 | 2.13e-01 | 2.69e-01 | 1.46e-05 | 5.93e-20 | δ2.2 |
| GPR171 | 2.43e-01 | 2.97e-01 | 1.8e-07 | 1.23e-19 | δ2.2 |
| TRAT1 | 2.84e-01 | 2.65e-01 | 6.42e-10 | 2.58e-17 | δ2.2 |
| GPR183 | 2.12e-01 | 2.68e-01 | 1.06e-04 | 4.16e-16 | δ2.2 |
| CD44 | 2.12e-01 | 2.21e-01 | 4.05e-07 | 3.49e-14 | δ2.2 |
| PDCD4 | 2.43e-01 | 2.05e-01 | 3.25e-08 | 1.96e-12 | δ2.2 |
| CCR6 | 5.42e-01 | 9.7e-01 | 2.25e-17 | 1.93e-73 | δ2.3 |
| SLC4A10 | 7.8e-01 | 8.52e-01 | 2.93e-34 | 1.12e-71 | δ2.3 |
| KLRB1 | 6.33e-01 | 7.79e-01 | 5.09e-17 | 1.46e-53 | δ2.3 |
| GZMK | 2.34e-01 | 6.22e-01 | 1.31e-09 | 3.77e-50 | δ2.3 |
| AQP3 | 5.42e-01 | 8.06e-01 | 3.04e-10 | 1.07e-37 | δ2.3 |
| NCR3 | 5.09e-01 | 8.58e-01 | 1.48e-08 | 7.39e-36 | δ2.3 |
| RPL10 | 2.03e-01 | 2.54e-01 | 9.3e-16 | 2.69e-33 | δ2.3 |
| LTB | 6.77e-01 | 7.17e-01 | 2.77e-20 | 3.05e-31 | δ2.3 |
| IL23R | 5.09e-01 | 4.68e-01 | 8.59e-16 | 1.43e-28 | δ2.3 |
| LST1 | 5.19e-01 | 6.74e-01 | 4.72e-11 | 2.75e-28 | δ2.3 |
| RPL13 | 2.48e-01 | 2.41e-01 | 1.83e-18 | 3.3e-28 | δ2.3 |
| RPS2 | 2.45e-01 | 2.46e-01 | 2.78e-18 | 5.6e-28 | δ2.3 |
| RPS18 | 2.41e-01 | 2.54e-01 | 1.92e-17 | 1.27e-26 | δ2.3 |
| RORC | 3.69e-01 | 4.17e-01 | 1.08e-12 | 4.63e-25 | δ2.3 |
| LTK | 4.69e-01 | 3.43e-01 | 2.62e-20 | 6.08e-25 | δ2.3 |
| RPL34 | 2.46e-01 | 2.32e-01 | 2.3e-19 | 5.13e-24 | δ2.3 |
| LGALS3 | 4.72e-01 | 4.24e-01 | 3.26e-12 | 4.61e-23 | δ2.3 |
| RPL32 | 2.18e-01 | 2.37e-01 | 1.87e-12 | 3.16e-22 | δ2.3 |

| | | | | | |
|---|---|---|---|---|---|
| S100A4 | 3.05e-01 | 3.9e-01 | 1.17e-11 | 4.34e-22 | δ2.3 |
| TMIGD2 | 2.5e-01 | 4.66e-01 | 6.36e-08 | 5.48e-22 | δ2.3 |
| EEF1A1 | 2.37e-01 | 2.12e-01 | 3.19e-17 | 1.41e-21 | δ2.3 |
| AMICA1 | 5.61e-01 | 5.9e-01 | 7.52e-11 | 4.42e-19 | δ2.3 |
| RPLP0 | 3.33e-01 | 3.37e-01 | 8.78e-14 | 3.32e-18 | δ2.3 |
| IL7R | 4.36e-01 | 4.02e-01 | 1.85e-11 | 3.67e-17 | δ2.3 |
| RPLP1 | 2.99e-01 | 2.27e-01 | 1.97e-15 | 2.49e-15 | δ2.3 |
| DPP4 | 4.43e-01 | 4.55e-01 | 2.91e-10 | 2.29e-14 | δ2.3 |
| CTSH | 2.16e-01 | 3.9e-01 | 1.14e-05 | 8.33e-14 | δ2.3 |
| RPS13 | 2.22e-01 | 2.2e-01 | 7.66e-10 | 8.42e-14 | δ2.3 |
| PHACTR2 | 3.66e-01 | 4.86e-01 | 2.92e-06 | 1.65e-13 | δ2.3 |
| ALOX5AP | 3.32e-01 | 4.36e-01 | 3.59e-08 | 2.37e-13 | δ2.3 |
| RPL12 | 2.22e-01 | 2.29e-01 | 2.61e-09 | 8.35e-13 | δ2.3 |
| CTSA | 2.65e-01 | 4.99e-01 | 1.95e-03 | 4.07e-12 | δ2.3 |
| GPR183 | 2.89e-01 | 4.77e-01 | 1.11e-03 | 1.04e-11 | δ2.3 |
| RPL10A | 2.27e-01 | 2.02e-01 | 7.88e-11 | 7.05e-11 | δ2.3 |
| TNFRSF25 | 3.77e-01 | 4.56e-01 | 1.08e-06 | 9.4e-11 | δ2.3 |
| CEBPD | 2.84e-01 | 4.71e-01 | 5.19e-03 | 5.07e-10 | δ2.3 |
| S100A6 | 3.06e-01 | 3.17e-01 | 5.37e-09 | 5.61e-10 | δ2.3 |
| GYG1 | 3.04e-01 | 4.04e-01 | 6.11e-04 | 1.94e-08 | δ2.3 |
| YWHAH | 2.46e-01 | 3.98e-01 | 1.71e-03 | 3.34e-08 | δ2.3 |
| GPR65 | 2.88e-01 | 4.17e-01 | 4.52e-04 | 4.24e-08 | δ2.3 |
| GPR171 | 2.83e-01 | 3.92e-01 | 3.53e-04 | 5.64e-08 | δ2.3 |
| TNFAIP8 | 2.48e-01 | 3.56e-01 | 2.51e-03 | 2.1e-07 | δ2.3 |
| CD28 | 2.6e-01 | 3.48e-01 | 1.56e-03 | 3.61e-07 | δ2.3 |
| MKNK1 | 2.89e-01 | 2.07e-01 | 1.58e-04 | 1.13e-06 | δ2.3 |
| NBEAL1 | 2.4e-01 | 2.82e-01 | 2.03e-04 | 2.28e-06 | δ2.3 |
| MAF | 3.17e-01 | 2.02e-01 | 9.26e-05 | 7.42e-06 | δ2.3 |
| CCR2 | 3.1e-01 | 2.42e-01 | 2.23e-05 | 1.43e-05 | δ2.3 |
| EDEM2 | 2.78e-01 | 2.17e-01 | 2.47e-03 | 2.01e-04 | δ2.3 |
| CCR5 | 2.27e-01 | 2.48e-01 | 2.42e-02 | 2.16e-03 | δ2.3 |
| SESN1 | 2.44e-01 | 2.02e-01 | 1.64e-02 | 3.2e-03 | δ2.3 |
| CMC1 | 3.15e-01 | 3.59e-01 | 5.31e-03 | 3.84e-03 | δ2.3 |

**Table A.2:** List of differentially expressed genes between δ1.1 and δ1.2 γδ-T cell subtypes from PBMC.

| gene | HD45 avg_logFC | HD6 avg_logFC | max pval | min pval | cluster |
|---|---|---|---|---|---|
| RPL21 | 3.55e-01 | 4.49e-01 | 1.67e-17 | 7.86e-81 | δ1.1 |
| RPS14 | 3.8e-01 | 5.19e-01 | 5e-17 | 2.27e-78 | δ1.1 |
| RPS29 | 2.87e-01 | 4.23e-01 | 7.51e-15 | 4.79e-68 | δ1.1 |
| RPL32 | 2.75e-01 | 3.48e-01 | 5.23e-12 | 5.66e-50 | δ1.1 |
| RPL34 | 3.07e-01 | 3.29e-01 | 3.32e-14 | 1.87e-49 | δ1.1 |
| RPLP2 | 3.6e-01 | 3.52e-01 | 3e-17 | 1.03e-48 | δ1.1 |
| RPS6 | 2.21e-01 | 3.36e-01 | 1.86e-07 | 2.13e-47 | δ1.1 |
| RPS27 | 2.23e-01 | 2.59e-01 | 1.11e-11 | 2.82e-46 | δ1.1 |
| RPS25 | 2.83e-01 | 3.73e-01 | 1.31e-07 | 9.94e-43 | δ1.1 |
| RPS28 | 2.5e-01 | 3.31e-01 | 1.64e-08 | 1.17e-41 | δ1.1 |
| RPS12 | 2.23e-01 | 3.56e-01 | 2.35e-05 | 1.14e-40 | δ1.1 |
| RPS27A | 1.78e-01 | 2.98e-01 | 4.99e-04 | 3.63e-40 | δ1.1 |
| RPL14 | 2.6e-01 | 4.73e-01 | 4.33e-06 | 2.16e-36 | δ1.1 |
| RPL31 | 1.97e-01 | 3.33e-01 | 2.17e-03 | 3.7e-35 | δ1.1 |
| RPL3 | 1.76e-01 | 2.67e-01 | 6.32e-05 | 6.27e-35 | δ1.1 |
| RPL27A | 2.81e-01 | 3.34e-01 | 2.03e-08 | 2.72e-34 | δ1.1 |
| RPS3 | 2.2e-01 | 2.94e-01 | 1.18e-06 | 3.23e-31 | δ1.1 |
| RPL35A | 2.11e-01 | 3.08e-01 | 3.89e-04 | 4.3e-31 | δ1.1 |
| RPL11 | 2e-01 | 2.78e-01 | 3.55e-05 | 4.18e-30 | δ1.1 |
| RPL30 | 2.87e-01 | 3.25e-01 | 1.58e-06 | 4.19e-30 | δ1.1 |
| RPL13 | 1.57e-01 | 2.39e-01 | 1.76e-04 | 2.65e-29 | δ1.1 |
| RPS3A | 1.65e-01 | 2.53e-01 | 5.72e-05 | 2.54e-26 | δ1.1 |
| EEF1A1 | 1.51e-01 | 2e-01 | 4.7e-04 | 4e-24 | δ1.1 |
| RPL38 | 3.85e-01 | 3.7e-01 | 2.79e-09 | 3.53e-23 | δ1.1 |
| RPL39 | 2.26e-01 | 2.46e-01 | 4.61e-06 | 2.03e-22 | δ1.1 |
| RPL7 | 1.15e-01 | 2.26e-01 | 1.28e-02 | 3.81e-21 | δ1.1 |
| RPL13A | 1.55e-01 | 1.89e-01 | 7.99e-05 | 5.14e-21 | δ1.1 |
| RPL9 | 2.28e-01 | 2.96e-01 | 1.04e-04 | 2.29e-20 | δ1.1 |
| RPL23A | 1.08e-01 | 2.19e-01 | 1.38e-02 | 7.38e-20 | δ1.1 |
| RPS13 | 1.28e-01 | 2.36e-01 | 1.89e-02 | 1.46e-19 | δ1.1 |
| RPS15 | 1.68e-01 | 2.36e-01 | 3.04e-03 | 4.31e-19 | δ1.1 |
| RPS19 | 1.96e-01 | 2.19e-01 | 4.27e-05 | 1e-18 | δ1.1 |
| RPL19 | 2.21e-01 | 2.14e-01 | 9.06e-07 | 1.04e-18 | δ1.1 |
| RPL41 | 1.08e-01 | 1.59e-01 | 3.83e-03 | 1.6e-16 | δ1.1 |

| | | | | | |
|---|---|---|---|---|---|
| RPS20 | 3.31e-01 | 2.63e-01 | 8.69e-07 | 2.31e-16 | δ1.1 |
| RPL36 | 2.8e-01 | 2.58e-01 | 7.61e-08 | 2.46e-16 | δ1.1 |
| RPS15A | 1.79e-01 | 1.89e-01 | 7.78e-05 | 3.5e-16 | δ1.1 |
| NELL2 | 5.16e-01 | 6.23e-01 | 1.14e-04 | 1.65e-15 | δ1.1 |
| RPLP1 | 2.4e-01 | 2.68e-01 | 6.99e-04 | 7.74e-15 | δ1.1 |
| RPS18 | 2.42e-01 | 1.75e-01 | 2e-09 | 1.48e-14 | δ1.1 |
| LDHB | 2.35e-01 | 4.05e-01 | 4.46e-02 | 9.08e-14 | δ1.1 |
| MT.ND3 | 1.22e-01 | 3.11e-01 | 6.78e-02 | 2.15e-13 | δ1.1 |
| RPS21 | 2.04e-01 | 2.72e-01 | 4.34e-03 | 3.19e-13 | δ1.1 |
| RPL35 | 3.52e-01 | 2.61e-01 | 8.65e-08 | 6.09e-13 | δ1.1 |
| EEF1B2 | 2.61e-01 | 3.23e-01 | 1.81e-03 | 1.4e-12 | δ1.1 |
| TPT1 | 2.28e-01 | 2.38e-01 | 8.31e-05 | 2.04e-12 | δ1.1 |
| CXCR4 | 4.37e-01 | 5.22e-01 | 2.32e-04 | 1.64e-11 | δ1.1 |
| CCR7 | 1.6e-01 | 6.09e-01 | 4.75e-01 | 1.91e-11 | δ1.1 |
| PIK3IP1 | 1.76e-01 | 5.54e-01 | 4.04e-02 | 4.95e-11 | δ1.1 |
| RPL5 | 1.29e-01 | 2.11e-01 | 1.6e-01 | 6.52e-10 | δ1.1 |
| RPS16 | 1.71e-01 | 1.92e-01 | 3.7e-03 | 8.86e-10 | δ1.1 |
| PFDN5 | 2.07e-01 | 3.08e-01 | 6.71e-03 | 6.39e-09 | δ1.1 |
| RPL10A | 1.37e-01 | 1.63e-01 | 2.68e-02 | 1.65e-08 | δ1.1 |
| RPL4 | 1.68e-01 | 1.97e-01 | 1.74e-02 | 1.76e-08 | δ1.1 |
| RPL8 | 2.1e-01 | 1.89e-01 | 9.19e-04 | 2.67e-08 | δ1.1 |
| GLTSCR2 | 1.83e-01 | 3.03e-01 | 6.7e-02 | 3.06e-08 | δ1.1 |
| IL7R | 4.99e-01 | 4.3e-01 | 3.44e-05 | 5.18e-08 | δ1.1 |
| RPL37 | 1.55e-01 | 1.73e-01 | 2.82e-02 | 8.6e-08 | δ1.1 |
| MT.CO3 | 1.42e-01 | 2.21e-01 | 6.07e-02 | 9.21e-08 | δ1.1 |
| RPL22 | 1.47e-01 | 1.97e-01 | 3.81e-02 | 3.02e-07 | δ1.1 |
| SELL | 2.45e-01 | 3.06e-01 | 1.29e-02 | 4.21e-07 | δ1.1 |
| RPS4X | 1.47e-01 | 1.28e-01 | 5.88e-04 | 6.07e-07 | δ1.1 |
| RPL36A | 2.87e-01 | 2.58e-01 | 3.86e-04 | 6.69e-07 | δ1.1 |
| RPL12 | 1.98e-01 | 1.3e-01 | 1.01e-03 | 1.47e-06 | δ1.1 |
| MT2A | 1.46e-01 | 4.48e-01 | 2.85e-01 | 3.66e-06 | δ1.1 |
| GIMAP4 | 1.51e-01 | 3.74e-01 | 1.33e-01 | 4.38e-06 | δ1.1 |
| SOX4 | 1.08e+00 | 1.44e+00 | 4.04e-11 | 9.52e-55 | δ1.2 |
| CD7 | 8.13e-01 | 8.54e-01 | 6.17e-14 | 2.29e-34 | δ1.2 |
| TMSB4X | 2.38e-01 | 2.81e-01 | 6.59e-08 | 1.08e-30 | δ1.2 |
| ZNF683 | 1.22e+00 | 1.32e+00 | 7.07e-11 | 3.58e-28 | δ1.2 |
| PFN1 | 3.21e-01 | 4.65e-01 | 1.91e-05 | 3.05e-25 | δ1.2 |
| ACTB | 2.99e-01 | 4.09e-01 | 5.13e-05 | 2.37e-23 | δ1.2 |

| | | | | | |
|---|---|---|---|---|---|
| SMC4 | 7.01e-01 | 1.04e+00 | 9.89e-05 | 6.03e-21 | δ1.2 |
| CHI3L2 | 4.93e-01 | 7.71e-01 | 1.47e-05 | 5.18e-19 | δ1.2 |
| PTPRC | 3.23e-01 | 4.94e-01 | 6.67e-04 | 1.54e-18 | δ1.2 |
| DOK2 | 5.06e-01 | 7.39e-01 | 1.46e-03 | 3.74e-18 | δ1.2 |
| PPP2R5C | 6.36e-01 | 6.15e-01 | 2.65e-06 | 4.08e-18 | δ1.2 |
| H3F3A | 2.65e-01 | 4.42e-01 | 6.66e-03 | 5.3e-18 | δ1.2 |
| MYL6 | 4.53e-01 | 4.54e-01 | 4.84e-07 | 7.07e-17 | δ1.2 |
| CHST2 | 3.74e-01 | 6.15e-01 | 2.24e-03 | 1.91e-15 | δ1.2 |
| CD3D | 3.46e-01 | 4.35e-01 | 7.23e-04 | 4.2e-15 | δ1.2 |
| CFL1 | 2.9e-01 | 3.53e-01 | 1.29e-05 | 4.84e-15 | δ1.2 |
| RPS24 | 2.03e-01 | 3.52e-01 | 1.25e-03 | 6.09e-15 | δ1.2 |
| SLC25A5 | 4.73e-01 | 5.74e-01 | 1.09e-03 | 7.82e-15 | δ1.2 |
| HIST1H2AC | 5.27e-01 | 6.64e-01 | 3.23e-04 | 1.29e-14 | δ1.2 |
| CLIC1 | 1.17e-01 | 5.87e-01 | 4.81e-01 | 5.85e-14 | δ1.2 |
| C16ORF54 | 2.74e-01 | 7.2e-01 | 6.17e-02 | 7.8e-14 | δ1.2 |
| ACTG1 | 2.88e-01 | 4.36e-01 | 2.46e-02 | 2.17e-13 | δ1.2 |
| CDC42 | 1.99e-01 | 4.66e-01 | 1.78e-01 | 2.65e-13 | δ1.2 |
| TOX | 3.01e-01 | 4.78e-01 | 7.46e-03 | 3.04e-13 | δ1.2 |
| SLFN5 | 1.34e-01 | 5.3e-01 | 2.5e-01 | 6.16e-13 | δ1.2 |
| STMN1 | 9.28e-01 | 5.6e-01 | 2.21e-07 | 1.75e-12 | δ1.2 |
| SET | 2.36e-01 | 5.59e-01 | 1.96e-01 | 1.97e-12 | δ1.2 |
| C4ORF48 | 2.16e-01 | 5.66e-01 | 2.04e-01 | 6.67e-12 | δ1.2 |
| ITGB2 | 1.89e-01 | 4.93e-01 | 2.45e-01 | 6.84e-12 | δ1.2 |
| LSP1 | 3.54e-01 | 4.69e-01 | 1.51e-04 | 7.01e-12 | δ1.2 |
| CD52 | 3.3e-01 | 3.49e-01 | 3.8e-04 | 1.81e-11 | δ1.2 |
| CTSW | 4.7e-01 | 4.94e-01 | 6.97e-04 | 2.01e-11 | δ1.2 |
| RAC2 | 2.48e-01 | 4.51e-01 | 6.05e-02 | 5.84e-11 | δ1.2 |
| ARPC3 | 2.06e-01 | 3.96e-01 | 6.63e-02 | 6.32e-11 | δ1.2 |
| SH3BGRL3 | 2.43e-01 | 3.87e-01 | 2.35e-02 | 9.14e-11 | δ1.2 |
| ARPC2 | 2.19e-01 | 3.75e-01 | 1.17e-02 | 2e-10 | δ1.2 |
| BAZ2B | 5.27e-01 | 4.39e-01 | 6.23e-04 | 5.75e-10 | δ1.2 |
| RRAS2 | 3.69e-01 | 4.07e-01 | 1.36e-03 | 7.09e-10 | δ1.2 |
| IL32 | 2.89e-01 | 5.15e-01 | 4.84e-04 | 9.73e-10 | δ1.2 |
| FAM173A | 4.2e-01 | 5.27e-01 | 7.8e-04 | 2.04e-09 | δ1.2 |
| EVL | 3.69e-01 | 3.45e-01 | 1.21e-03 | 4.07e-09 | δ1.2 |
| DBN1 | 2.38e-01 | 3.87e-01 | 3e-03 | 4.71e-09 | δ1.2 |
| CXXC5 | 5.62e-01 | 3.01e-01 | 2.94e-04 | 9.61e-09 | δ1.2 |
| MATK | 3.85e-01 | 4.85e-01 | 6.02e-02 | 1.76e-08 | δ1.2 |

| | | | | | |
|---|---|---|---|---|---|
| PRKCH | 4.93e-01 | 4.39e-01 | 4.94e-05 | 2.23e-08 | δ1.2 |
| RHOH | 5.36e-01 | 4.45e-01 | 1.74e-04 | 2.75e-08 | δ1.2 |
| PTPRCAP | 1.54e-01 | 2.97e-01 | 8.02e-02 | 3.64e-08 | δ1.2 |
| IFITM2 | 3.11e-01 | 4.69e-01 | 1.84e-02 | 4.64e-08 | δ1.2 |
| EIF5A | 2.84e-01 | 3.5e-01 | 1.47e-01 | 6.01e-08 | δ1.2 |
| HMGN1 | 3.94e-01 | 3.88e-01 | 2.37e-04 | 7.14e-08 | δ1.2 |
| EMP3 | 1.3e-01 | 4.1e-01 | 2.84e-01 | 7.7e-08 | δ1.2 |
| IP6K2 | 4.57e-01 | 4.66e-01 | 1.51e-03 | 8.49e-08 | δ1.2 |
| MBNL1 | 2.07e-01 | 4.48e-01 | 3.56e-01 | 9.88e-08 | δ1.2 |
| PLAC8 | 1.31e-01 | 4.54e-01 | 7.52e-01 | 1.05e-07 | δ1.2 |
| DDAH2 | 2.9e-01 | 4.07e-01 | 1.32e-02 | 1.16e-07 | δ1.2 |
| C12ORF75 | 2.79e-01 | 4.64e-01 | 3.42e-03 | 1.36e-07 | δ1.2 |
| PSME2 | 2.79e-01 | 4.18e-01 | 5.51e-02 | 1.76e-07 | δ1.2 |
| DYNLL1 | 3.26e-01 | 4.48e-01 | 6.6e-02 | 2.07e-07 | δ1.2 |
| CD3G | 3.56e-01 | 3.65e-01 | 8.17e-03 | 2.18e-07 | δ1.2 |
| LEF1 | 3.57e-01 | 3.53e-01 | 5.46e-03 | 2.2e-07 | δ1.2 |
| HMGN2 | 3.31e-01 | 2.97e-01 | 3.18e-02 | 3.11e-07 | δ1.2 |
| UBB | 4.06e-01 | 3.04e-01 | 5.64e-05 | 3.47e-07 | δ1.2 |
| YBX1 | 1.54e-01 | 2.6e-01 | 2.84e-02 | 4.47e-07 | δ1.2 |
| CCDC167 | 2.49e-01 | 4.44e-01 | 1.4e-01 | 5.29e-07 | δ1.2 |
| AOAH | 2.93e-01 | 4.1e-01 | 5.74e-02 | 5.49e-07 | δ1.2 |
| IFITM1 | 3.4e-01 | 4.11e-01 | 2.25e-03 | 5.78e-07 | δ1.2 |
| MARCKSL1 | 4.83e-01 | 4.71e-01 | 3e-03 | 1.05e-06 | δ1.2 |
| HIST1H2BK | 2.97e-01 | 3.83e-01 | 5.86e-03 | 1.45e-06 | δ1.2 |
| ANXA6 | 2.91e-01 | 3.84e-01 | 6.64e-02 | 1.65e-06 | δ1.2 |
| FCGRT | 3.63e-01 | 3.82e-01 | 5.51e-03 | 1.73e-06 | δ1.2 |
| ATP5B | 2.41e-01 | 3.7e-01 | 1.21e-01 | 1.93e-06 | δ1.2 |
| BIN2 | 3.79e-01 | 3.45e-01 | 1.73e-03 | 2.22e-06 | δ1.2 |
| CALM2 | 1.69e-01 | 3.03e-01 | 5.47e-02 | 2.54e-06 | δ1.2 |
| FNBP1 | 1.05e-01 | 4.04e-01 | 7.75e-01 | 2.6e-06 | δ1.2 |
| PSME1 | 2.63e-01 | 3.25e-01 | 3.22e-02 | 3.11e-06 | δ1.2 |
| TBC1D10C | 1.06e-01 | 3.38e-01 | 6.96e-01 | 4.33e-06 | δ1.2 |
| COX6C | 2.43e-01 | 2.76e-01 | 9.37e-02 | 6.07e-06 | δ1.2 |
| SSBP4 | 1.77e-01 | 3.73e-01 | 2.01e-01 | 8.39e-06 | δ1.2 |
| HMGB1 | 1.77e-01 | 2.59e-01 | 5.28e-02 | 8.6e-06 | δ1.2 |
| PTP4A2 | 3.26e-01 | 3.72e-01 | 3.09e-02 | 8.74e-06 | δ1.2 |
| PARP1 | 3.72e-01 | 3.85e-01 | 6.85e-03 | 8.83e-06 | δ1.2 |
| PRMT1 | 5.74e-01 | 3.61e-01 | 1.26e-04 | 9.07e-06 | δ1.2 |

| | | | | | |
|---|---|---|---|---|---|
| MYOM2 | 5.39e-01 | 3.11e-01 | 7.66e-02 | 1.1e-05 | δ1.2 |
| TCF7 | 2.23e-01 | 3.16e-01 | 1.3e-02 | 1.13e-05 | δ1.2 |
| IKZF2 | 3.53e-01 | 3.64e-01 | 2.18e-02 | 1.22e-05 | δ1.2 |
| UCP2 | 2.3e-01 | 3.83e-01 | 2e-02 | 1.45e-05 | δ1.2 |
| RPIA | 3.66e-01 | 4.3e-01 | 3.29e-03 | 1.48e-05 | δ1.2 |

**Table A.3:** List of differentially expressed genes between δ2.2 and δ2.3 γδ-T cell subtypes from PBMC.

| gene | HD45 avg_logFC | HD6 avg_logFC | max pval | min pval | cluster |
|---|---|---|---|---|---|
| CCL5 | 5.79e-01 | 5.11e-01 | 2.46e-33 | 1.34e-41 | δ2.2 |
| KLRD1 | 6.25e-01 | 1.02e+00 | 3.51e-14 | 7.25e-34 | δ2.2 |
| XCL1 | 5.08e-01 | 9.31e-01 | 3.07e-07 | 8.75e-26 | δ2.2 |
| CST7 | 3.95e-01 | 5.53e-01 | 3.46e-09 | 4.17e-24 | δ2.2 |
| MALAT1 | 1.84e-01 | 1.61e-01 | 9.22e-11 | 1.43e-21 | δ2.2 |
| NKG7 | 1.94e-01 | 4.14e-01 | 3.14e-05 | 3.69e-19 | δ2.2 |
| KLRC1 | 8.99e-01 | 9.26e-01 | 3.65e-14 | 1.87e-18 | δ2.2 |
| GZMH | 2.67e-01 | 8.73e-01 | 2.93e-02 | 1.85e-17 | δ2.2 |
| GZMB | 3.7e-01 | 7.75e-01 | 3.03e-03 | 2.73e-14 | δ2.2 |
| CCL4 | 3.87e-01 | 5.75e-01 | 1.15e-03 | 2.4e-11 | δ2.2 |
| CD74 | 2.38e-01 | 4.83e-01 | 2.74e-03 | 4.56e-11 | δ2.2 |
| GNLY | 4.79e-01 | 9.42e-01 | 1.6e-02 | 5.6e-11 | δ2.2 |
| SELL | 2.61e-01 | 4.93e-01 | 9.12e-03 | 1.6e-09 | δ2.2 |
| HLA.DPA1 | 3.71e-01 | 5.04e-01 | 2.64e-04 | 1.27e-08 | δ2.2 |
| HLA.DPB1 | 4.54e-01 | 4.96e-01 | 2.09e-05 | 1.36e-08 | δ2.2 |
| TPST2 | 2.31e-01 | 4.86e-01 | 1.73e-02 | 1.83e-08 | δ2.2 |
| DUSP2 | 6.39e-01 | 4.21e-01 | 1.37e-07 | 2.35e-08 | δ2.2 |
| PTPRC | 2.47e-01 | 2.54e-01 | 2.84e-05 | 6.43e-08 | δ2.2 |
| HLA.DRB1 | 6.01e-01 | 4.67e-01 | 9.61e-07 | 1.64e-07 | δ2.2 |
| CD300A | 4.2e-01 | 4.54e-01 | 3.7e-05 | 1.67e-07 | δ2.2 |
| LITAF | 4.03e-01 | 3.83e-01 | 6.76e-06 | 3.38e-07 | δ2.2 |
| GZMM | 2.31e-01 | 3.54e-01 | 8.47e-03 | 1.07e-06 | δ2.2 |
| GZMK | 2.44e-01 | 1.44e-01 | 2.39e-05 | 1.37e-06 | δ2.2 |
| C1ORF21 | 2.53e-01 | 3.83e-01 | 9.81e-03 | 1.45e-06 | δ2.2 |
| PYHIN1 | 2.45e-01 | 4.21e-01 | 1.1e-02 | 1.64e-06 | δ2.2 |
| HLA.DRB5 | 3.86e-01 | 3.31e-01 | 6.94e-06 | 3.37e-06 | δ2.2 |
| FTH1 | 1.37e-01 | 2.42e-01 | 1.04e-01 | 3.85e-06 | δ2.2 |
| CLIC3 | 2.34e-01 | 4.36e-01 | 4.42e-02 | 4.32e-06 | δ2.2 |
| HLA.DMA | 3.16e-01 | 3.09e-01 | 2.53e-04 | 8.11e-06 | δ2.2 |
| FGR | 3.75e-01 | 3.36e-01 | 5.83e-05 | 1.45e-05 | δ2.2 |
| SYTL3 | 3.09e-01 | 3.92e-01 | 1.76e-03 | 1.56e-05 | δ2.2 |
| SLC4A10 | 7.14e-01 | 8.22e-01 | 9.8e-23 | 8.1e-47 | δ2.3 |
| CCR6 | 4.65e-01 | 9.06e-01 | 2.71e-11 | 1.05e-44 | δ2.3 |
| RPL10 | 1.39e-01 | 2.13e-01 | 1.3e-08 | 3.49e-29 | δ2.3 |

| | | | | | |
|---|---|---|---|---|---|
| TMIGD2 | 2.62e-01 | 5.12e-01 | 5.46e-08 | 3.63e-27 | δ2.3 |
| KLRB1 | 3.35e-01 | 4.85e-01 | 1.36e-05 | 4.85e-27 | δ2.3 |
| RPL34 | 1.76e-01 | 2.09e-01 | 3.62e-12 | 8.72e-25 | δ2.3 |
| NCR3 | 4.16e-01 | 7.28e-01 | 1.99e-05 | 2.93e-24 | δ2.3 |
| LTB | 5.38e-01 | 6.84e-01 | 6.69e-12 | 3.34e-24 | δ2.3 |
| S100A4 | 3.45e-01 | 4e-01 | 1.78e-13 | 2.67e-23 | δ2.3 |
| AQP3 | 4.16e-01 | 6.98e-01 | 9.04e-06 | 9.82e-23 | δ2.3 |
| RPL32 | 1.55e-01 | 2.11e-01 | 5.89e-07 | 1.23e-22 | δ2.3 |
| LST1 | 4.87e-01 | 6.26e-01 | 1.31e-08 | 1.29e-19 | δ2.3 |
| IL23R | 4.57e-01 | 4.45e-01 | 2.76e-10 | 2.49e-17 | δ2.3 |
| RPL13 | 1.47e-01 | 1.56e-01 | 4.99e-09 | 3.98e-17 | δ2.3 |
| EEF1A1 | 1.87e-01 | 1.6e-01 | 1.21e-12 | 2.4e-16 | δ2.3 |
| LTK | 4.49e-01 | 3.38e-01 | 2.63e-14 | 2.66e-16 | δ2.3 |
| RORC | 3.51e-01 | 3.93e-01 | 1.8e-09 | 5.9e-15 | δ2.3 |
| LGALS3 | 4.12e-01 | 4e-01 | 1.93e-07 | 1.14e-14 | δ2.3 |
| RPS13 | 1.82e-01 | 2.2e-01 | 7.22e-07 | 1.51e-14 | δ2.3 |
| RPL18A | 1.22e-01 | 1.81e-01 | 7.25e-05 | 2.32e-14 | δ2.3 |
| RPS2 | 1.57e-01 | 1.52e-01 | 3.01e-09 | 4.82e-13 | δ2.3 |
| RPS18 | 1.27e-01 | 1.46e-01 | 1.78e-06 | 3.37e-12 | δ2.3 |
| AMICA1 | 4.55e-01 | 4.85e-01 | 2.11e-06 | 5.26e-12 | δ2.3 |
| CTSH | 2.04e-01 | 3.78e-01 | 3.39e-05 | 1.5e-10 | δ2.3 |
| RPS19 | 1.21e-01 | 1.38e-01 | 2.79e-04 | 1.44e-09 | δ2.3 |
| CTSA | 2.21e-01 | 4.51e-01 | 5.19e-03 | 1.08e-08 | δ2.3 |
| S100A6 | 3.21e-01 | 3.01e-01 | 8.06e-09 | 1.17e-08 | δ2.3 |
| RPL11 | 1.48e-01 | 1.48e-01 | 3.03e-07 | 1.44e-08 | δ2.3 |
| RPL8 | 1.14e-01 | 1.83e-01 | 8.58e-04 | 1.68e-08 | δ2.3 |
| RPL29 | 1.33e-01 | 2.03e-01 | 1.24e-02 | 7.92e-08 | δ2.3 |
| RPS5 | 1.24e-01 | 1.95e-01 | 8.23e-03 | 1.77e-07 | δ2.3 |
| EOMES | 1.57e-01 | 3.8e-01 | 2.64e-02 | 2.88e-07 | δ2.3 |
| IL7R | 1.94e-01 | 2.38e-01 | 1.89e-03 | 4.32e-07 | δ2.3 |
| TNFRSF25 | 3.16e-01 | 3.86e-01 | 1.11e-04 | 6.09e-07 | δ2.3 |
| YWHAH | 2.6e-01 | 3.81e-01 | 1.8e-03 | 7.82e-07 | δ2.3 |
| RPLP0 | 2.25e-01 | 1.8e-01 | 7.25e-06 | 1.21e-06 | δ2.3 |
| GPR65 | 2.23e-01 | 3.8e-01 | 6.33e-03 | 1.22e-06 | δ2.3 |
| PTGER2 | 2.11e-01 | 3.44e-01 | 1.58e-03 | 1.47e-06 | δ2.3 |
| RPL12 | 1.19e-01 | 1.55e-01 | 1.2e-03 | 1.64e-06 | δ2.3 |
| RPL10A | 1.49e-01 | 1.41e-01 | 8.82e-06 | 5.37e-06 | δ2.3 |
| ALOX5AP | 1.81e-01 | 2.92e-01 | 8.69e-04 | 5.99e-06 | δ2.3 |

| | | | | | |
|---|---|---|---|---|---|
| CAMK4 | 1.58e-01 | 3.18e-01 | 2.67e-01 | 9.63e-06 | δ2.3 |

**Table A.4:** List of differentially expressed genes between γδ-T cell subtypes in breast tumour samples (BC1 and BC2). Associated with Figure X.

| gene | BC1 avg_logFC | BC2 avg_logFC | max pval | min pval | cluster |
|---|---|---|---|---|---|
| FGFBP2 | 2.04e+00 | 1.85e+00 | 2.04e-20 | 1.72e-86 | γδ-T.1 |
| RAP1GAP2 | 1.72e+00 | 1.37e+00 | 6.33e-05 | 2.35e-52 | γδ-T.1 |
| BC043356 | 1.32e+00 | 8.21e-01 | 6.99e-03 | 1.32e-44 | γδ-T.1 |
| GNLY1 | 1.46e+00 | 2.23e+00 | 7.53e-12 | 1.56e-36 | γδ-T.1 |
| SPON2 | 1.07e+00 | 1.58e+00 | 4.82e-07 | 4.12e-33 | γδ-T.1 |
| PLAC8 | 1.14e+00 | 1.35e+00 | 1.1e-06 | 8.31e-27 | γδ-T.1 |
| GZMB1 | 1.17e+00 | 1.97e+00 | 5.97e-08 | 3.33e-25 | γδ-T.1 |
| GZMH | 1.06e+00 | 1.27e+00 | 2.06e-07 | 1.29e-23 | γδ-T.1 |
| LINC00299 | 1.28e+00 | 1.26e+00 | 4.88e-04 | 6.11e-23 | γδ-T.1 |
| UBE2F | 1.11e+00 | 9.15e-01 | 4.58e-04 | 3.3e-19 | γδ-T.1 |
| KLRF1 | 1.13e+00 | 1.9e+00 | 5.78e-09 | 7.38e-18 | γδ-T.1 |
| GPR56 | 5.64e-01 | 4.28e-01 | 2.28e-02 | 1.61e-17 | γδ-T.1 |
| PDGFD1 | 1.39e+00 | 1.85e+00 | 4.45e-08 | 1.92e-17 | γδ-T.1 |
| KLRD11 | 8.65e-01 | 1.52e+00 | 5.26e-07 | 5.21e-17 | γδ-T.1 |
| BNC2 | 1.1e+00 | 1.29e+00 | 6.63e-07 | 1.15e-16 | γδ-T.1 |
| FAM53B | 9.37e-01 | 2.55e-01 | 2.52e-01 | 2.23e-16 | γδ-T.1 |
| FCGR3A | 6.2e-01 | 7.64e-01 | 1.21e-04 | 5.92e-16 | γδ-T.1 |
| TGFBR31 | 1e+00 | 9.78e-01 | 3.97e-03 | 2.86e-15 | γδ-T.1 |
| ZEB21 | 7.37e-01 | 1.34e+00 | 2.31e-05 | 8.84e-15 | γδ-T.1 |
| RAP2B | 8.22e-01 | 6.45e-01 | 1.33e-01 | 5.33e-14 | γδ-T.1 |
| TTC38 | 7.48e-01 | 8.74e-01 | 8.72e-05 | 1.19e-13 | γδ-T.1 |
| PTPN12 | 9.16e-01 | 1.97e+00 | 1.19e-11 | 1.29e-13 | γδ-T.1 |
| KLHDC4 | 7.14e-01 | 4.3e-01 | 4.66e-01 | 2.62e-12 | γδ-T.1 |
| BCR | 7.11e-01 | 1.34e+00 | 4.54e-05 | 4.52e-12 | γδ-T.1 |
| PLEKHF1 | 6.51e-01 | 6.25e-01 | 1.48e-02 | 8.68e-12 | γδ-T.1 |
| GK5 | 8.35e-01 | 9.58e-01 | 5.15e-04 | 1.07e-11 | γδ-T.1 |
| ITGB21 | 6.58e-01 | 6.07e-01 | 1.12e-01 | 6.66e-11 | γδ-T.1 |
| CX3CR1 | 4.68e-01 | 4.86e-01 | 2.03e-02 | 7.09e-11 | γδ-T.1 |
| DGKD | 7.75e-01 | 4.26e-01 | 3.49e-02 | 1.18e-10 | γδ-T.1 |
| ARPC2 | 5.1e-01 | 4.37e-01 | 1.82e-01 | 2.97e-10 | γδ-T.1 |
| MYBL11 | 5.9e-01 | 7.93e-01 | 3.48e-03 | 3.61e-10 | γδ-T.1 |
| PTGDS | 5.54e-01 | 1.26e+00 | 8.67e-05 | 3.86e-10 | γδ-T.1 |
| SLCO4C1 | 4.43e-01 | 6.32e-01 | 1.64e-05 | 1.39e-09 | γδ-T.1 |
| MYL12A | 5.5e-01 | 9.19e-01 | 8.17e-05 | 1.51e-09 | γδ-T.1 |

| | | | | | |
|---|---|---|---|---|---|
| ABI3 | 6.7e-01 | 9.78e-01 | 2.44e-02 | 1.68e-09 | γδ-T.1 |
| C12ORF75 | 6.39e-01 | 5.96e-01 | 1.3e-03 | 1.9e-09 | γδ-T.1 |
| TYROBP1 | 7.1e-01 | 1.35e+00 | 4.4e-06 | 1.97e-09 | γδ-T.1 |
| DOCK51 | 1.02e+00 | 2.79e-01 | 1.49e-01 | 2.01e-09 | γδ-T.1 |
| RASA3 | 7.52e-01 | 8.89e-01 | 9.18e-03 | 3.63e-09 | γδ-T.1 |
| HCST1 | 5.61e-01 | 7.89e-01 | 2.37e-02 | 3.97e-09 | γδ-T.1 |
| PTPN7 | 7.59e-01 | 4.4e-01 | 1.46e-01 | 4.21e-09 | γδ-T.1 |
| PXN | 5.66e-01 | 3.15e-01 | 1.84e-01 | 6.03e-09 | γδ-T.1 |
| MYO1F1 | 7.32e-01 | 1.45e+00 | 1.68e-06 | 6.1e-09 | γδ-T.1 |
| B3GNT71 | 7.43e-01 | 4.81e-01 | 4.87e-03 | 6.58e-09 | γδ-T.1 |
| AKR1C3 | 4.04e-01 | 7.33e-01 | 5.73e-05 | 7.51e-09 | γδ-T.1 |
| ITGAL | 7.43e-01 | 5.48e-01 | 3.68e-02 | 7.92e-09 | γδ-T.1 |
| SYTL31 | 5.34e-01 | 1.44e+00 | 5.02e-06 | 8.26e-09 | γδ-T.1 |
| HAVCR21 | 7.56e-01 | 1.15e+00 | 9.18e-03 | 1.82e-08 | γδ-T.1 |
| SSBP3 | 7.13e-01 | 3.11e-01 | 5.99e-01 | 2.94e-08 | γδ-T.1 |
| BIN21 | 6.95e-01 | 6.2e-01 | 3.44e-02 | 4.82e-08 | γδ-T.1 |
| NFE2L21 | 6.4e-01 | 5.05e-01 | 1.65e-01 | 4.83e-08 | γδ-T.1 |
| AOAH1 | 4.99e-01 | 1.29e+00 | 1.26e-05 | 5.66e-08 | γδ-T.1 |
| FOXK2 | 7.68e-01 | 6.3e-01 | 3.95e-01 | 7.92e-08 | γδ-T.1 |
| DSTN | 6.47e-01 | 6.1e-01 | 6.7e-03 | 8.16e-08 | γδ-T.1 |
| LINGO21 | 1.43e+00 | 1.88e+00 | 2.61e-03 | 1.62e-07 | γδ-T.1 |
| ACTB | 5.19e-01 | 7.92e-01 | 6.93e-04 | 1.67e-07 | γδ-T.1 |
| PFN1 | 5.34e-01 | 6.49e-01 | 2.11e-02 | 1.83e-07 | γδ-T.1 |
| LPCAT1 | 6.01e-01 | 7.77e-01 | 5.66e-03 | 2.13e-07 | γδ-T.1 |
| CCL4 | 5.89e-01 | 1.84e+00 | 1.75e-04 | 3.27e-07 | γδ-T.1 |
| TNFRSF1B | 5.65e-01 | 3.27e-01 | 6.58e-02 | 5.64e-07 | γδ-T.1 |
| GZMA1 | 2.65e-01 | 1.52e+00 | 1.08e-04 | 5.72e-07 | γδ-T.1 |
| ITGAM | 3.98e-01 | 6.78e-01 | 3.56e-02 | 6.06e-07 | γδ-T.1 |
| PLEKHA21 | 5.69e-01 | 3.89e-01 | 6.34e-02 | 6.31e-07 | γδ-T.1 |
| TMEM181 | 5.27e-01 | 8.88e-01 | 2.69e-02 | 6.63e-07 | γδ-T.1 |
| TSPAN32 | 4.8e-01 | 8.66e-01 | 6.37e-03 | 7.09e-07 | γδ-T.1 |
| PTPRE | 5.48e-01 | 5.22e-01 | 3.1e-01 | 7.67e-07 | γδ-T.1 |
| ST3GAL4 | 4.08e-01 | 3.6e-01 | 3.47e-02 | 8.03e-07 | γδ-T.1 |
| UPP1 | 5.16e-01 | 4.11e-01 | 2.27e-02 | 8.95e-07 | γδ-T.1 |
| SETBP1 | 6.69e-01 | 6.04e-01 | 1.39e-01 | 1e-06 | γδ-T.1 |
| DDIT4 | 6.81e-01 | 3.4e-01 | 4.64e-01 | 1.1e-06 | γδ-T.1 |
| S100A4 | 4.21e-01 | 7.37e-01 | 5.4e-03 | 1.34e-06 | γδ-T.1 |
| SYNE1 | 5.95e-01 | 9.99e-01 | 1.27e-02 | 1.52e-06 | γδ-T.1 |

| | | | | | |
|---|---|---|---|---|---|
| CCND3 | 4.57e-01 | 3.37e-01 | 2.89e-01 | 1.64e-06 | γδ-T.1 |
| BHLHE401 | 5.63e-01 | 7.19e-01 | 3.35e-02 | 1.94e-06 | γδ-T.1 |
| DIP2A | 5.63e-01 | 1.11e+00 | 4.06e-04 | 2.85e-06 | γδ-T.1 |
| HLA.C | 3.53e-01 | 7.54e-01 | 1.48e-03 | 3.22e-06 | γδ-T.1 |
| CLIC3 | 2.59e-01 | 9.91e-01 | 2.32e-03 | 3.25e-06 | γδ-T.1 |
| SAMD31 | 3.85e-01 | 1.4e+00 | 9.82e-05 | 4.02e-06 | γδ-T.1 |
| PLCB11 | 2.67e-01 | 1.39e+00 | 9.7e-04 | 5.58e-06 | γδ-T.1 |
| TRAPPC10 | 4.76e-01 | 7.27e-01 | 7.42e-02 | 6.49e-06 | γδ-T.1 |
| RNF168 | 4.66e-01 | 1.11e+00 | 2.56e-04 | 9.11e-06 | γδ-T.1 |
| FNDC3B1 | 6.65e-01 | 7.43e-01 | 1.16e-02 | 1.02e-05 | γδ-T.1 |
| HIPK2 | 6.66e-01 | 6.73e-01 | 1.04e-01 | 1.04e-05 | γδ-T.1 |
| ARL4C | 4.53e-01 | 8.8e-01 | 2.31e-02 | 1.23e-05 | γδ-T.1 |
| SPN | 5.18e-01 | 8.99e-01 | 2.93e-03 | 1.3e-05 | γδ-T.1 |
| ST8SIA6 | 3.62e-01 | 4.07e-01 | 1.92e-02 | 1.31e-05 | γδ-T.1 |
| RHBDF2 | 4.23e-01 | 3.29e-01 | 6.42e-02 | 1.36e-05 | γδ-T.1 |
| EMP3 | 4.47e-01 | 9.16e-01 | 1.66e-05 | 1.42e-05 | γδ-T.1 |
| CD38 | 4.63e-01 | 1.71e+00 | 5.9e-05 | 1.46e-05 | γδ-T.1 |
| ACTG1 | 4.73e-01 | 4.5e-01 | 2.02e-01 | 1.82e-05 | γδ-T.1 |
| STK38 | 5.35e-01 | 7.23e-01 | 5.48e-02 | 1.84e-05 | γδ-T.1 |
| CMIP1 | 4.92e-01 | 9.12e-01 | 1.49e-02 | 2.42e-05 | γδ-T.1 |
| PAFAH2 | 2.9e-01 | 7.4e-01 | 3.67e-03 | 2.44e-05 | γδ-T.1 |
| IGF1R | 8.57e-01 | 8.51e-01 | 6.79e-02 | 2.64e-05 | γδ-T.1 |
| GFOD11 | 4.29e-01 | 1.27e+00 | 3.71e-03 | 2.66e-05 | γδ-T.1 |
| VAV3 | 4.96e-01 | 9.81e-01 | 2.03e-02 | 2.66e-05 | γδ-T.1 |
| VPS37B | 4.78e-01 | 4.18e-01 | 6.23e-01 | 2.76e-05 | γδ-T.1 |
| CCND2 | 3.47e-01 | 1.18e+00 | 1.46e-03 | 2.8e-05 | γδ-T.1 |
| IL21R | 4.16e-01 | 2.74e-01 | 3.81e-01 | 3.22e-05 | γδ-T.1 |
| FGR | 3.02e-01 | 7.06e-01 | 9.55e-05 | 4.68e-05 | γδ-T.1 |
| NFATC2 | 3.5e-01 | 9.35e-01 | 3.35e-04 | 4.82e-05 | γδ-T.1 |
| SLA21 | 5.93e-01 | 8.07e-01 | 1.45e-02 | 4.89e-05 | γδ-T.1 |
| AK5 | 3.6e-01 | 8.13e-01 | 1.72e-04 | 5.49e-05 | γδ-T.1 |
| F2R | 3.94e-01 | 5.93e-01 | 4.65e-03 | 5.64e-05 | γδ-T.1 |
| ADRB21 | 4.32e-01 | 4.69e-01 | 1.99e-01 | 5.68e-05 | γδ-T.1 |
| GZMM1 | 3.73e-01 | 9.38e-01 | 8.41e-04 | 5.7e-05 | γδ-T.1 |
| SLC39A10 | 3.1e-01 | 7.58e-01 | 2.54e-02 | 6.15e-05 | γδ-T.1 |
| SUSD1 | 3.46e-01 | 9.2e-01 | 9.76e-04 | 6.42e-05 | γδ-T.1 |
| LINC00861 | 4.14e-01 | 8e-01 | 9.52e-02 | 6.57e-05 | γδ-T.1 |
| CASP10 | 4.34e-01 | 3.59e-01 | 3.52e-01 | 6.82e-05 | γδ-T.1 |

| | | | | | |
|---|---|---|---|---|---|
| SH3GLB1 | 3.09e-01 | 3.01e-01 | 4.34e-01 | 7.29e-05 | γδ-T.1 |
| CCDC88C | 3.62e-01 | 4.37e-01 | 3.25e-01 | 7.48e-05 | γδ-T.1 |
| SUN2 | 3.61e-01 | 5.3e-01 | 1.97e-01 | 7.94e-05 | γδ-T.1 |
| LOC284757 | 5.47e-01 | 1.15e+00 | 1e-02 | 9.36e-05 | γδ-T.1 |
| GNAL | 5.92e-01 | 5.96e-01 | 1.26e-01 | 1.02e-04 | γδ-T.1 |
| GSTP11 | 3.95e-01 | 5.04e-01 | 1.07e-01 | 1.46e-04 | γδ-T.1 |
| KLRB11 | 6.92e-01 | 1.27e+00 | 3.82e-04 | 1.49e-04 | γδ-T.1 |
| PITPNC1 | 3.52e-01 | 1.36e+00 | 1.43e-03 | 1.5e-04 | γδ-T.1 |
| AXIN1 | 5.95e-01 | 8.86e-01 | 4.7e-03 | 1.54e-04 | γδ-T.1 |
| LLGL2 | 4.29e-01 | 3.67e-01 | 1.42e-01 | 1.55e-04 | γδ-T.1 |
| SLC1A4 | 6.23e-01 | 5.85e-01 | 1.32e-01 | 1.79e-04 | γδ-T.1 |
| BCL2 | 3.95e-01 | 2.88e-01 | 5.06e-01 | 1.96e-04 | γδ-T.1 |
| YAF2 | 5.02e-01 | 8.13e-01 | 3.38e-02 | 1.98e-04 | γδ-T.1 |
| HOPX1 | 3.54e-01 | 1.43e+00 | 1.32e-03 | 2.16e-04 | γδ-T.1 |
| PHF20 | 4.27e-01 | 3.41e-01 | 5.55e-01 | 2.36e-04 | γδ-T.1 |
| SSX2IP | 2.64e-01 | 8.55e-01 | 7.78e-04 | 2.4e-04 | γδ-T.1 |
| GSAP | 4.6e-01 | 9.5e-01 | 1.4e-02 | 2.47e-04 | γδ-T.1 |
| TES | 3.8e-01 | 5.51e-01 | 7.69e-02 | 2.64e-04 | γδ-T.1 |
| BPGM | 4.89e-01 | 7.82e-01 | 4.89e-02 | 2.76e-04 | γδ-T.1 |
| HSPA5 | 3.35e-01 | 5.58e-01 | 5.25e-03 | 2.79e-04 | γδ-T.1 |
| RALGAPA1 | 3.69e-01 | 7.92e-01 | 8.28e-03 | 2.83e-04 | γδ-T.1 |
| ITGA5 | 2.92e-01 | 7.33e-01 | 4.93e-03 | 3.08e-04 | γδ-T.1 |
| TMEM50B.1 | 3.92e-01 | 3.8e-01 | 5.78e-02 | 3.1e-04 | γδ-T.1 |
| PTGER2 | 3.5e-01 | 5.31e-01 | 1.4e-01 | 3.11e-04 | γδ-T.1 |
| PRR5L1 | 4.71e-01 | 8.32e-01 | 8.32e-02 | 3.72e-04 | γδ-T.1 |
| GNG2 | 4.04e-01 | 1.08e+00 | 3.39e-03 | 3.89e-04 | γδ-T.1 |
| MCTP21 | 4.09e-01 | 1.11e+00 | 8.98e-04 | 4.6e-04 | γδ-T.1 |
| B4GALT51 | 2.59e-01 | 4.71e-01 | 4.33e-01 | 4.67e-04 | γδ-T.1 |
| PPP2R5C | 2.7e-01 | 7.44e-01 | 5.61e-03 | 4.69e-04 | γδ-T.1 |
| PIK3CD | 3.78e-01 | 6.38e-01 | 1.65e-01 | 4.75e-04 | γδ-T.1 |
| PCSK7 | 3.2e-01 | 6.62e-01 | 1.06e-02 | 5.16e-04 | γδ-T.1 |
| PPM1L | 5.65e-01 | 8.89e-01 | 8.27e-04 | 5.55e-04 | γδ-T.1 |
| CDC42SE1 | 6.1e-01 | 7.62e-01 | 4.43e-02 | 6.14e-04 | γδ-T.1 |
| GNPTAB | 4.86e-01 | 4.27e-01 | 1.16e-01 | 7.4e-04 | γδ-T.1 |
| STARD3NL1 | 2.97e-01 | 8.57e-01 | 2.08e-02 | 8.88e-04 | γδ-T.1 |
| EFHD2 | 4.24e-01 | 7.16e-01 | 1.17e-03 | 1.02e-03 | γδ-T.1 |
| CYBA | 2.98e-01 | 5.42e-01 | 1.2e-02 | 1.02e-03 | γδ-T.1 |
| NAPA | 3.48e-01 | 3.55e-01 | 4.61e-01 | 1.05e-03 | γδ-T.1 |

| | | | | | |
|---|---|---|---|---|---|
| HCP5 | 3.02e-01 | 9.13e-01 | 4.88e-03 | 1.06e-03 | γδ-T.1 |
| FAM49B | 3.53e-01 | 9.32e-01 | 5.6e-03 | 1.06e-03 | γδ-T.1 |
| ARSG | 4.18e-01 | 1.24e+00 | 1.35e-03 | 1.11e-03 | γδ-T.1 |
| BRD7 | 3.89e-01 | 5.28e-01 | 2.89e-01 | 1.24e-03 | γδ-T.1 |
| MSL2 | 3.7e-01 | 5.59e-01 | 9.86e-02 | 1.27e-03 | γδ-T.1 |
| CEP78 | 4.05e-01 | 8.37e-01 | 1.01e-03 | 1.49e-03 | γδ-T.1 |
| NDUFA12 | 3.67e-01 | 4.43e-01 | 1.94e-01 | 1.52e-03 | γδ-T.1 |
| RNF167 | 3.13e-01 | 4.47e-01 | 3.16e-01 | 1.63e-03 | γδ-T.1 |
| FOSL21 | 3.6e-01 | 3.86e-01 | 5.3e-01 | 1.83e-03 | γδ-T.1 |
| CNOT21 | 4.51e-01 | 7.78e-01 | 3.07e-01 | 2.01e-03 | γδ-T.1 |
| CARD11 | 4.27e-01 | 7.47e-01 | 2.26e-01 | 2.16e-03 | γδ-T.1 |
| ST3GAL5 | 3.5e-01 | 4.17e-01 | 3.93e-01 | 2.24e-03 | γδ-T.1 |
| RUNX3 | 2.99e-01 | 8.89e-01 | 1.41e-01 | 2.29e-03 | γδ-T.1 |
| TBCB | 3.76e-01 | 3.68e-01 | 1.39e-01 | 2.35e-03 | γδ-T.1 |
| C5ORF56 | 2.99e-01 | 5.39e-01 | 2.13e-01 | 2.4e-03 | γδ-T.1 |
| ST3GAL2 | 3.72e-01 | 3.26e-01 | 2.44e-01 | 2.5e-03 | γδ-T.1 |
| AES | 3.2e-01 | 9.71e-01 | 2.47e-02 | 2.79e-03 | γδ-T.1 |
| ADAM10 | 4.55e-01 | 4.63e-01 | 3.97e-01 | 2.86e-03 | γδ-T.1 |
| ZBTB2 | 3.66e-01 | 3.52e-01 | 4.56e-01 | 2.89e-03 | γδ-T.1 |
| IFITM11 | 3.01e-01 | 8.21e-01 | 5.91e-03 | 2.97e-03 | γδ-T.1 |
| GPR651 | 2.88e-01 | 1.14e+00 | 1.12e-02 | 3.22e-03 | γδ-T.1 |
| OSTF1 | 2.97e-01 | 4.58e-01 | 4.06e-02 | 3.33e-03 | γδ-T.1 |
| CALM1 | 2.58e-01 | 2.97e-01 | 8.97e-02 | 3.35e-03 | γδ-T.1 |
| ESYT2 | 3.94e-01 | 5.04e-01 | 9.51e-02 | 3.65e-03 | γδ-T.1 |
| JAK1 | 2.57e-01 | 3.03e-01 | 3e-01 | 3.8e-03 | γδ-T.1 |
| UBN2 | 3.37e-01 | 5.83e-01 | 1e-01 | 3.83e-03 | γδ-T.1 |
| LNPEP | 3.27e-01 | 3.86e-01 | 3.97e-01 | 4.03e-03 | γδ-T.1 |
| ABL1 | 4.13e-01 | 7.52e-01 | 6.64e-02 | 4.07e-03 | γδ-T.1 |
| PDE7A | 2.57e-01 | 5.22e-01 | 6.69e-01 | 4.13e-03 | γδ-T.1 |
| DDOST | 3.18e-01 | 4.97e-01 | 1.31e-02 | 4.38e-03 | γδ-T.1 |
| C1ORF56 | 6.17e-01 | 7.11e-01 | 2.51e-02 | 4.51e-03 | γδ-T.1 |
| GTF3C1 | 3.48e-01 | 4.56e-01 | 2.03e-01 | 4.72e-03 | γδ-T.1 |
| JAZF1 | 3.7e-01 | 2.53e-01 | 6.49e-01 | 4.79e-03 | γδ-T.1 |
| SAE1 | 3.4e-01 | 4.69e-01 | 2.11e-01 | 4.85e-03 | γδ-T.1 |
| CFL1 | 2.98e-01 | 6.38e-01 | 6.07e-03 | 5.77e-03 | γδ-T.1 |
| RBM22 | 2.93e-01 | 2.81e-01 | 5.39e-01 | 5.98e-03 | γδ-T.1 |
| FANCA | 2.73e-01 | 5.42e-01 | 1.93e-01 | 5.98e-03 | γδ-T.1 |
| AES.1 | 3.38e-01 | 9.11e-01 | 1.8e-02 | 6.02e-03 | γδ-T.1 |

| | | | | | |
|---|---|---|---|---|---|
| RAP1B | 2.63e-01 | 2.93e-01 | 5.89e-01 | 6.08e-03 | γδ-T.1 |
| HIRA | 3.49e-01 | 3.26e-01 | 5.07e-01 | 6.09e-03 | γδ-T.1 |
| PHAX | 3.04e-01 | 4.64e-01 | 2.11e-01 | 6.9e-03 | γδ-T.1 |
| REV3L | 3.21e-01 | 4.85e-01 | 2e-01 | 7.4e-03 | γδ-T.1 |
| WDR37 | 4.06e-01 | 7.28e-01 | 7.39e-02 | 8.03e-03 | γδ-T.1 |
| VTI1B | 2.91e-01 | 2.89e-01 | 2.08e-01 | 8.24e-03 | γδ-T.1 |
| LAT | 3.76e-01 | 5.1e-01 | 4.05e-01 | 8.35e-03 | γδ-T.1 |
| AP2M1 | 3.21e-01 | 2.57e-01 | 5.12e-01 | 8.66e-03 | γδ-T.1 |
| HSH2D | 3.72e-01 | 4.52e-01 | 2.62e-01 | 8.66e-03 | γδ-T.1 |
| OXSR1 | 3.08e-01 | 5.01e-01 | 3.36e-01 | 8.94e-03 | γδ-T.1 |
| SYAP1 | 3.85e-01 | 3.95e-01 | 1.77e-01 | 9.31e-03 | γδ-T.1 |
| ARRB1 | 3.14e-01 | 3.29e-01 | 9.83e-02 | 9.83e-03 | γδ-T.1 |
| TMEM71 | 5.11e-01 | 9.13e-01 | 8.89e-03 | 1.08e-02 | γδ-T.1 |
| LPP | 2.71e-01 | 5.19e-01 | 4.77e-02 | 1.34e-02 | γδ-T.1 |
| LDHA | 3.17e-01 | 2.77e-01 | 1.69e-01 | 1.37e-02 | γδ-T.1 |
| EIF4G3 | 3.61e-01 | 4.09e-01 | 3.02e-01 | 1.53e-02 | γδ-T.1 |
| TRAPPC8 | 2.63e-01 | 3.31e-01 | 6.02e-01 | 1.59e-02 | γδ-T.1 |
| GAB3 | 4.08e-01 | 3.72e-01 | 3.09e-02 | 1.74e-02 | γδ-T.1 |
| MED14 | 2.71e-01 | 4.35e-01 | 1.05e-01 | 1.85e-02 | γδ-T.1 |
| PPP3R1 | 2.8e-01 | 7.92e-01 | 2.32e-02 | 1.96e-02 | γδ-T.1 |
| JA700183 | 2.75e-01 | 4.74e-01 | 3.31e-01 | 1.98e-02 | γδ-T.1 |
| FKBP111 | 2.79e-01 | 5.79e-01 | 1.62e-02 | 2.01e-02 | γδ-T.1 |
| CCL31 | 5.23e-01 | 8.41e-01 | 9.35e-02 | 2.05e-02 | γδ-T.1 |
| ABHD21 | 3.31e-01 | 5.17e-01 | 3.33e-02 | 2.1e-02 | γδ-T.1 |
| KANSL1 | 2.72e-01 | 4.51e-01 | 2.32e-01 | 2.22e-02 | γδ-T.1 |
| NPIPL1.1 | 2.71e-01 | 3.48e-01 | 1.54e-01 | 2.29e-02 | γδ-T.1 |
| PTPN41 | 2.62e-01 | 5.99e-01 | 6.02e-02 | 2.34e-02 | γδ-T.1 |
| ITCH | 2.72e-01 | 8.06e-01 | 1.87e-02 | 2.42e-02 | γδ-T.1 |
| CD226 | 2.54e-01 | 9.45e-01 | 5.23e-02 | 2.45e-02 | γδ-T.1 |
| DOK6 | 2.85e-01 | 5.44e-01 | 2.41e-01 | 2.61e-02 | γδ-T.1 |
| RIN31 | 2.8e-01 | 5.9e-01 | 5.14e-02 | 2.64e-02 | γδ-T.1 |
| UCK2 | 2.73e-01 | 2.65e-01 | 1.53e-01 | 2.65e-02 | γδ-T.1 |
| IL2RG | 2.88e-01 | 3.8e-01 | 2.68e-01 | 2.76e-02 | γδ-T.1 |
| TG1 | 2.73e-01 | 8.06e-01 | 3.23e-02 | 2.81e-02 | γδ-T.1 |
| MANF | 3.31e-01 | 4.17e-01 | 3.51e-01 | 2.84e-02 | γδ-T.1 |
| UBR1 | 3.32e-01 | 4.39e-01 | 3.7e-01 | 2.88e-02 | γδ-T.1 |
| STRN | 3.56e-01 | 6.54e-01 | 3.78e-02 | 2.93e-02 | γδ-T.1 |
| PAPD5 | 3.66e-01 | 8.45e-01 | 2.1e-02 | 3.09e-02 | γδ-T.1 |

| | | | | | |
|---|---|---|---|---|---|
| MECP2 | 3.21e-01 | 3.84e-01 | 2.39e-01 | 3.19e-02 | γδ-T.1 |
| MTF2 | 3.1e-01 | 6.58e-01 | 2.03e-01 | 3.26e-02 | γδ-T.1 |
| RIPK2 | 2.6e-01 | 4.17e-01 | 3.88e-01 | 3.33e-02 | γδ-T.1 |
| PCED1B | 2.81e-01 | 5.65e-01 | 3.88e-02 | 3.34e-02 | γδ-T.1 |
| SIRT2 | 3.26e-01 | 2.92e-01 | 6.09e-01 | 3.43e-02 | γδ-T.1 |
| GRAP2 | 2.65e-01 | 5.76e-01 | 9.8e-02 | 3.69e-02 | γδ-T.1 |
| TARS | 3.21e-01 | 5.25e-01 | 1.66e-01 | 3.72e-02 | γδ-T.1 |
| MYO1G | 2.84e-01 | 4.76e-01 | 2.28e-01 | 3.76e-02 | γδ-T.1 |
| RAD9A | 2.59e-01 | 3.12e-01 | 6.06e-01 | 3.94e-02 | γδ-T.1 |
| JMJD6 | 2.81e-01 | 5.15e-01 | 6.9e-02 | 4.27e-02 | γδ-T.1 |
| SMAD5 | 2.51e-01 | 7.76e-01 | 2.97e-02 | 4.4e-02 | γδ-T.1 |
| KDM2A | 2.5e-01 | 5.98e-01 | 5.77e-02 | 4.42e-02 | γδ-T.1 |
| C18ORF8 | 2.85e-01 | 4.28e-01 | 1.01e-01 | 4.45e-02 | γδ-T.1 |
| SECISBP2L | 2.8e-01 | 3.06e-01 | 5.47e-02 | 4.69e-02 | γδ-T.1 |
| ETFA | 3.11e-01 | 6.03e-01 | 3.89e-01 | 4.75e-02 | γδ-T.1 |
| POMP | 2.67e-01 | 4.56e-01 | 5.95e-02 | 4.9e-02 | γδ-T.1 |
| SET | 2.63e-01 | 3.83e-01 | 3.21e-01 | 4.9e-02 | γδ-T.1 |
| MED15 | 2.62e-01 | 4.32e-01 | 5.94e-02 | 5.18e-02 | γδ-T.1 |
| C2CD5 | 3.21e-01 | 4.02e-01 | 4.85e-02 | 5.28e-02 | γδ-T.1 |
| IRF2BPL | 3.16e-01 | 3.4e-01 | 3.31e-01 | 5.75e-02 | γδ-T.1 |
| DDX52 | 2.78e-01 | 4.43e-01 | 4.08e-01 | 6.07e-02 | γδ-T.1 |
| LUZP6 | 2.83e-01 | 6.79e-01 | 1.24e-01 | 6.4e-02 | γδ-T.1 |
| GLCCI1 | 3.51e-01 | 4e-01 | 2.88e-01 | 6.55e-02 | γδ-T.1 |
| USP28 | 2.6e-01 | 4.44e-01 | 4.29e-02 | 7.04e-02 | γδ-T.1 |
| KIAA2018 | 2.69e-01 | 3.24e-01 | 4.15e-01 | 7.2e-02 | γδ-T.1 |
| PSEN1 | 3.17e-01 | 2.81e-01 | 3.63e-01 | 7.43e-02 | γδ-T.1 |
| MADD | 3.05e-01 | 3.75e-01 | 4.82e-01 | 7.46e-02 | γδ-T.1 |
| CHMP1B | 3.34e-01 | 4.14e-01 | 3.34e-01 | 7.61e-02 | γδ-T.1 |
| PDIA6 | 2.64e-01 | 4.58e-01 | 2.43e-01 | 7.7e-02 | γδ-T.1 |
| FAM65B | 3.15e-01 | 3.2e-01 | 3.4e-01 | 7.82e-02 | γδ-T.1 |
| PPP1R18 | 2.59e-01 | 6.24e-01 | 8.57e-02 | 7.9e-02 | γδ-T.1 |
| AK097472 | 2.64e-01 | 3.49e-01 | 2.38e-01 | 8.23e-02 | γδ-T.1 |
| MAZ | 2.57e-01 | 3.42e-01 | 2.54e-01 | 9.79e-02 | γδ-T.1 |
| CPQ | 3.23e-01 | 4.91e-01 | 1.76e-01 | 9.91e-02 | γδ-T.1 |
| CDK6 | 2.58e-01 | 8.25e-01 | 2.08e-01 | 1e-01 | γδ-T.1 |
| THOC7 | 2.58e-01 | 3.21e-01 | 2.23e-01 | 1.06e-01 | γδ-T.1 |
| RBBP4 | 2.64e-01 | 3.11e-01 | 2.81e-01 | 1.19e-01 | γδ-T.1 |
| ZNF710 | 3.1e-01 | 7.43e-01 | 7.72e-02 | 1.2e-01 | γδ-T.1 |

| | | | | | |
|---|---|---|---|---|---|
| SLC12A2 | 2.53e-01 | 5.43e-01 | 3.39e-01 | 1.22e-01 | γδ-T.1 |
| SLK | 2.65e-01 | 6.7e-01 | 1.07e-01 | 1.23e-01 | γδ-T.1 |
| RAB18 | 2.6e-01 | 3.18e-01 | 6.78e-01 | 1.33e-01 | γδ-T.1 |
| KPNA3 | 2.74e-01 | 6.19e-01 | 1.22e-01 | 1.34e-01 | γδ-T.1 |
| PACSIN21 | 2.76e-01 | 3.72e-01 | 5.12e-01 | 1.47e-01 | γδ-T.1 |
| PAK1 | 3.21e-01 | 6.06e-01 | 2.42e-01 | 1.74e-01 | γδ-T.1 |
| RNF115 | 2.59e-01 | 3.9e-01 | 6.2e-01 | 1.91e-01 | γδ-T.1 |
| PPP2R3A | 2.85e-01 | 5.93e-01 | 1.09e-01 | 1.95e-01 | γδ-T.1 |
| C16ORF54 | 3.04e-01 | 3.71e-01 | 4.48e-01 | 2e-01 | γδ-T.1 |
| CPEB4 | 2.77e-01 | 3.13e-01 | 5.98e-01 | 2.96e-01 | γδ-T.1 |
| GNLY | 3.16e+00 | 3.2e+00 | 4.13e-90 | 6.23e-246 | γδ-T.2 |
| ATP8B4 | 2.02e+00 | 1.49e+00 | 2.68e-26 | 2.14e-177 | γδ-T.2 |
| NCAM1 | 1.97e+00 | 1.39e+00 | 8.63e-27 | 2.16e-163 | γδ-T.2 |
| TYROBP | 1.41e+00 | 1.99e+00 | 5.27e-55 | 2.69e-130 | γδ-T.2 |
| PLCG2 | 2.02e+00 | 5.93e-01 | 1.8e-04 | 1.07e-123 | γδ-T.2 |
| LOC285972 | 1.26e+00 | 7.94e-01 | 5.33e-08 | 2.78e-106 | γδ-T.2 |
| KLRD1 | 1.45e+00 | 1.44e+00 | 1.01e-25 | 2.13e-101 | γδ-T.2 |
| AV4S11 | 1.42e+00 | 1.95e+00 | 4.09e-41 | 2.23e-90 | γδ-T.2 |
| RIN3 | 1.35e+00 | 1.21e+00 | 1.3e-15 | 6.64e-73 | γδ-T.2 |
| KLRC1 | 1.2e+00 | 1.35e+00 | 2.89e-25 | 1.51e-72 | γδ-T.2 |
| GZMB | 1.68e+00 | 1.69e+00 | 6.35e-24 | 9.31e-70 | γδ-T.2 |
| FCER1G | 9.3e-01 | 1.14e+00 | 6.53e-28 | 6.87e-68 | γδ-T.2 |
| CLNK | 1e+00 | 4.53e-01 | 5.32e-03 | 1.17e-61 | γδ-T.2 |
| MCTP2 | 1.15e+00 | 1e+00 | 1.97e-12 | 3.25e-61 | γδ-T.2 |
| LINGO2 | 1.88e+00 | 1.08e+00 | 6.05e-11 | 4.47e-60 | γδ-T.2 |
| HAVCR2 | 1.25e+00 | 7.36e-01 | 1.26e-06 | 9.06e-60 | γδ-T.2 |
| CD7 | 1.04e+00 | 1.5e+00 | 3.39e-26 | 3.75e-57 | γδ-T.2 |
| B3GNT7 | 7.5e-01 | 1.56e+00 | 5.29e-27 | 4.14e-53 | γδ-T.2 |
| CNOT2 | 1.15e+00 | 8.66e-01 | 9.03e-12 | 1.12e-50 | γδ-T.2 |
| CD63 | 9.39e-01 | 1.24e+00 | 9.69e-19 | 2.1e-50 | γδ-T.2 |
| AK096314 | 8.39e-01 | 1.72e+00 | 3.3e-37 | 1.89e-49 | γδ-T.2 |
| KLRK1 | 8.22e-01 | 1.67e+00 | 1.21e-32 | 2.26e-48 | γδ-T.2 |
| DOCK5 | 1.1e+00 | 9.88e-01 | 3.87e-12 | 8.43e-48 | γδ-T.2 |
| WIPF3 | 6.41e-01 | 2.28e+00 | 1.41e-33 | 3.39e-46 | γδ-T.2 |
| GAS7 | 1.1e+00 | 1.34e+00 | 9.4e-15 | 1.36e-44 | γδ-T.2 |
| APBA2 | 1.03e+00 | 3.28e-01 | 1.81e-02 | 2.59e-42 | γδ-T.2 |
| PRKX | 8.38e-01 | 8.5e-01 | 4.34e-09 | 4.85e-42 | γδ-T.2 |
| PDGFD | 1.06e+00 | 1.17e+00 | 5.62e-11 | 1.45e-39 | γδ-T.2 |

| | | | | | |
|---|---|---|---|---|---|
| CCL5 | 5.67e-01 | 1.79e+00 | 4.51e-27 | 2.07e-39 | γδ-T.2 |
| ABCB11 | 9.09e-01 | 9.44e-01 | 4.29e-09 | 5.07e-39 | γδ-T.2 |
| SAMD3 | 8.4e-01 | 1.08e+00 | 2.53e-13 | 1.04e-38 | γδ-T.2 |
| HOPX | 9.69e-01 | 1.37e+00 | 7.35e-24 | 2.94e-38 | γδ-T.2 |
| KRT81 | 4.67e-01 | 1.19e+00 | 6.45e-22 | 5.21e-38 | γδ-T.2 |
| GEM | 1.12e+00 | 5.04e-01 | 5.92e-03 | 1.5e-37 | γδ-T.2 |
| GALNT2 | 9.14e-01 | 1.36e+00 | 7.56e-19 | 1.63e-37 | γδ-T.2 |
| PPP1R9A | 7.12e-01 | 8.01e-01 | 1.83e-06 | 3.94e-37 | γδ-T.2 |
| KRT86 | 5.12e-01 | 1.78e+00 | 5.55e-33 | 8.78e-36 | γδ-T.2 |
| SRGAP3 | 8.84e-01 | 1.06e+00 | 7.1e-11 | 9.44e-34 | γδ-T.2 |
| ABHD2 | 8.46e-01 | 7.15e-01 | 7.61e-08 | 1.51e-33 | γδ-T.2 |
| AHI1 | 8.92e-01 | 8.33e-01 | 9.21e-07 | 8.9e-33 | γδ-T.2 |
| ITGB1 | 7.9e-01 | 8.66e-01 | 1.33e-08 | 2.09e-32 | γδ-T.2 |
| AOAH | 4.19e-01 | 1.62e+00 | 1.58e-19 | 1.22e-31 | γδ-T.2 |
| LOC100506776 | 7.94e-01 | 6.4e-01 | 2.83e-05 | 6.17e-31 | γδ-T.2 |
| ZNF683 | 7.92e-01 | 1.01e+00 | 1.76e-11 | 2.57e-30 | γδ-T.2 |
| LAT2 | 5.96e-01 | 6.88e-01 | 2.02e-04 | 5.05e-29 | γδ-T.2 |
| SYTL3 | 3.64e-01 | 1.31e+00 | 2.83e-09 | 5.47e-29 | γδ-T.2 |
| LDB2 | 7.59e-01 | 1.08e+00 | 3.2e-16 | 4.87e-28 | γδ-T.2 |
| IFITM2 | 7.06e-01 | 6.64e-01 | 9.64e-06 | 1.16e-27 | γδ-T.2 |
| GZMA | 5.42e-01 | 1.63e+00 | 1.53e-12 | 1.92e-27 | γδ-T.2 |
| ITGA1 | 7.6e-01 | 1.73e+00 | 4.42e-27 | 1.97e-27 | γδ-T.2 |
| GPR97 | 4.5e-01 | 9.41e-01 | 1.44e-17 | 8.92e-26 | γδ-T.2 |
| L3MBTL4 | 8e-01 | 4.36e-01 | 2.62e-02 | 3.1e-24 | γδ-T.2 |
| DFNB31I | 6.75e-01 | 8.71e-01 | 1.69e-09 | 4.75e-23 | γδ-T.2 |
| KCNQ5 | 9.67e-01 | 2.84e-01 | 3.62e-02 | 1.55e-21 | γδ-T.2 |
| CBLB | 4.29e-01 | 1.23e+00 | 2.85e-16 | 3.56e-21 | γδ-T.2 |
| CMC1 | 9.59e-01 | 1.61e+00 | 9.03e-19 | 7.01e-20 | γδ-T.2 |
| PRR5L | 3.89e-01 | 1.38e+00 | 1.68e-04 | 2.42e-19 | γδ-T.2 |
| GFOD1 | 6.96e-01 | 6.19e-01 | 1.16e-03 | 3.35e-19 | γδ-T.2 |
| CMIP | 4.91e-01 | 1e+00 | 4.66e-17 | 8.53e-19 | γδ-T.2 |
| TNFRSF18 | 5.86e-01 | 1.37e+00 | 1.53e-16 | 1.03e-17 | γδ-T.2 |
| PLXNA4 | 6.2e-01 | 8.69e-01 | 5.01e-11 | 1.45e-17 | γδ-T.2 |
| EPHA41 | 6.11e-01 | 8.24e-01 | 2.05e-05 | 8.71e-17 | γδ-T.2 |
| GAB1 | 5.31e-01 | 5.92e-01 | 6.74e-04 | 2.78e-16 | γδ-T.2 |
| CHST12 | 5.5e-01 | 9.84e-01 | 2.3e-09 | 2.8e-16 | γδ-T.2 |
| ZEB2 | 4.75e-01 | 7.25e-01 | 3.37e-07 | 3.61e-16 | γδ-T.2 |
| IL2RB | 5.13e-01 | 6.84e-01 | 2.58e-06 | 8.12e-16 | γδ-T.2 |

| | | | | | |
|---|---|---|---|---|---|
| LRMP | 6.34e-01 | 5.95e-01 | 3.33e-03 | 4.5e-15 | γδ-T.2 |
| HCST | 2.85e-01 | 1.01e+00 | 5.5e-06 | 4.68e-15 | γδ-T.2 |
| MAFF1 | 2.9e-01 | 1.12e+00 | 1.59e-05 | 5.89e-15 | γδ-T.2 |
| LYST | 5.26e-01 | 2.88e-01 | 2.01e-01 | 1.18e-14 | γδ-T.2 |
| ATP8A1 | 4.41e-01 | 8.05e-01 | 4.42e-08 | 1.21e-14 | γδ-T.2 |
| PELO | 4.67e-01 | 7.77e-01 | 2.02e-08 | 1.73e-14 | γδ-T.2 |
| NCR31 | 3.34e-01 | 3.1e-01 | 3.28e-02 | 2.48e-12 | γδ-T.2 |
| MAFF.1 | 2.5e-01 | 8.58e-01 | 6.31e-06 | 2.84e-12 | γδ-T.2 |
| CHN2 | 5.02e-01 | 1.06e+00 | 3.02e-11 | 3.46e-12 | γδ-T.2 |
| HIP1 | 5.6e-01 | 7.2e-01 | 1.5e-04 | 1.43e-11 | γδ-T.2 |
| CLEC2B | 2.53e-01 | 1.04e+00 | 3.65e-05 | 3.92e-11 | γδ-T.2 |
| RBPJ | 2.66e-01 | 1.03e+00 | 2.21e-02 | 4.08e-11 | γδ-T.2 |
| MCF2L2 | 2.95e-01 | 1.18e+00 | 1.92e-03 | 5.69e-11 | γδ-T.2 |
| ZNF611 | 4.76e-01 | 3.86e-01 | 3.48e-02 | 8.44e-11 | γδ-T.2 |
| ITGAE | 2.57e-01 | 1.01e+00 | 5.07e-03 | 1.74e-10 | γδ-T.2 |
| ABTB2 | 8.52e-01 | 6.85e-01 | 2.35e-05 | 1.99e-10 | γδ-T.2 |
| FOSL2 | 4.84e-01 | 5.56e-01 | 2.65e-04 | 2.79e-10 | γδ-T.2 |
| LITAF1 | 4.82e-01 | 7.09e-01 | 1.2e-06 | 3.5e-10 | γδ-T.2 |
| SLA2 | 3.64e-01 | 9.96e-01 | 3.92e-06 | 5.01e-10 | γδ-T.2 |
| MRPS6 | 2.58e-01 | 8.5e-01 | 1.32e-04 | 6.98e-10 | γδ-T.2 |
| DOCK8 | 3.41e-01 | 5.93e-01 | 5.87e-07 | 9.11e-10 | γδ-T.2 |
| TFDP2 | 4.4e-01 | 3.08e-01 | 4.41e-01 | 2.01e-09 | γδ-T.2 |
| TCTN3 | 4.96e-01 | 4.38e-01 | 2.64e-02 | 2.77e-09 | γδ-T.2 |
| IFITM1 | 3.66e-01 | 5.81e-01 | 1.48e-06 | 3.95e-09 | γδ-T.2 |
| PARP8 | 2.99e-01 | 7.8e-01 | 4.52e-07 | 4.34e-09 | γδ-T.2 |
| PACSIN2 | 3.68e-01 | 3.4e-01 | 4.84e-02 | 6.38e-09 | γδ-T.2 |
| TIGIT | 3.65e-01 | 6.08e-01 | 1.04e-04 | 7.23e-09 | γδ-T.2 |
| SLC16A3 | 3.92e-01 | 7.26e-01 | 4.48e-07 | 9.16e-09 | γδ-T.2 |
| CCL3 | 3.45e-01 | 1.36e+00 | 1.89e-03 | 1.04e-08 | γδ-T.2 |
| PTPN4 | 2.64e-01 | 7.99e-01 | 2.8e-04 | 1.26e-08 | γδ-T.2 |
| CAPN12 | 3.58e-01 | 3.53e-01 | 2.14e-02 | 3.36e-08 | γδ-T.2 |
| FNDC3B | 3.32e-01 | 8.92e-01 | 6.6e-05 | 3.51e-08 | γδ-T.2 |
| BHLHE40 | 2.93e-01 | 8.12e-01 | 5.83e-05 | 4.22e-08 | γδ-T.2 |
| GSTP1 | 3.65e-01 | 5.86e-01 | 5.84e-04 | 4.54e-08 | γδ-T.2 |
| ERGIC1 | 3.82e-01 | 5.85e-01 | 5.73e-04 | 5.79e-08 | γδ-T.2 |
| LGALS1 | 3.35e-01 | 1.11e+00 | 9.03e-06 | 5.98e-08 | γδ-T.2 |
| CTNNB1 | 4.11e-01 | 3.57e-01 | 4.35e-03 | 6.46e-08 | γδ-T.2 |
| TPST2 | 3.56e-01 | 2.79e-01 | 1.38e-01 | 8.14e-08 | γδ-T.2 |

| | | | | | |
|---|---|---|---|---|---|
| PTPRJ | 2.62e-01 | 3.46e-01 | 1.02e-04 | 8.97e-08 | γδ-T.2 |
| PLEKHA2 | 3.72e-01 | 4.22e-01 | 7.34e-03 | 9.33e-08 | γδ-T.2 |
| BIN2 | 4.1e-01 | 4.05e-01 | 1.62e-02 | 9.51e-08 | γδ-T.2 |
| PIK3R1 | 3.73e-01 | 8e-01 | 2.38e-07 | 1.3e-07 | γδ-T.2 |
| SLAMF7 | 3.07e-01 | 2.66e-01 | 1.58e-02 | 1.92e-07 | γδ-T.2 |
| CD971 | 3.19e-01 | 7.93e-01 | 7.3e-05 | 3.05e-07 | γδ-T.2 |
| SOS2 | 4.27e-01 | 4.53e-01 | 2.33e-03 | 3.8e-07 | γδ-T.2 |
| CASK | 3.25e-01 | 2.74e-01 | 8.37e-02 | 4.34e-07 | γδ-T.2 |
| CTSD | 2.57e-01 | 7.91e-01 | 4.76e-06 | 4.58e-07 | γδ-T.2 |
| GALNT111 | 2.61e-01 | 7.17e-01 | 2.61e-04 | 5.02e-07 | γδ-T.2 |
| TMX4 | 3.95e-01 | 5.38e-01 | 1.08e-02 | 6.04e-07 | γδ-T.2 |
| SKI | 3.21e-01 | 6.96e-01 | 3.78e-06 | 6.58e-07 | γδ-T.2 |
| SFMBT2 | 2.92e-01 | 7.64e-01 | 7.27e-05 | 6.73e-07 | γδ-T.2 |
| C20ORF112 | 3.44e-01 | 8.26e-01 | 9.57e-07 | 7.4e-07 | γδ-T.2 |
| ID2 | 2.85e-01 | 9.32e-01 | 2.81e-04 | 7.99e-07 | γδ-T.2 |
| MNAT1 | 3.16e-01 | 2.97e-01 | 7.18e-02 | 9.49e-07 | γδ-T.2 |
| SLFN5 | 3.69e-01 | 4.37e-01 | 5.19e-03 | 2.06e-06 | γδ-T.2 |
| SACM1L | 2.66e-01 | 5.04e-01 | 2.51e-03 | 2.2e-06 | γδ-T.2 |
| B3GNT5 | 3.14e-01 | 8.42e-01 | 3.85e-04 | 2.38e-06 | γδ-T.2 |
| BRF1 | 3.12e-01 | 3.54e-01 | 4.38e-02 | 3.43e-06 | γδ-T.2 |
| TG | 2.53e-01 | 7.15e-01 | 1.57e-04 | 5.14e-06 | γδ-T.2 |
| SPECC1 | 3.41e-01 | 4.15e-01 | 7.56e-03 | 6.9e-06 | γδ-T.2 |
| PON2 | 3.21e-01 | 6.73e-01 | 2.34e-04 | 7.18e-06 | γδ-T.2 |
| MPG | 3.09e-01 | 4.57e-01 | 9.15e-03 | 7.47e-06 | γδ-T.2 |
| LY6E | 3.34e-01 | 4.5e-01 | 2.77e-02 | 7.68e-06 | γδ-T.2 |
| KDM5B | 3.21e-01 | 5.86e-01 | 2.91e-03 | 1.75e-05 | γδ-T.2 |
| PYHIN1 | 2.67e-01 | 4.51e-01 | 1.19e-02 | 1.9e-05 | γδ-T.2 |
| BTN3A1 | 2.72e-01 | 3.12e-01 | 3.26e-03 | 2.49e-05 | γδ-T.2 |
| STARD9 | 2.67e-01 | 4.33e-01 | 4.66e-03 | 3.35e-05 | γδ-T.2 |
| ZNF827 | 3.46e-01 | 2.67e-01 | 1.48e-01 | 3.87e-05 | γδ-T.2 |
| CRIM1 | 2.67e-01 | 7.12e-01 | 6.48e-05 | 5.23e-05 | γδ-T.2 |
| MIPEPP3 | 2.83e-01 | 3.38e-01 | 5.46e-02 | 7.3e-05 | γδ-T.2 |
| MYO1F | 3.25e-01 | 4.27e-01 | 2.58e-02 | 7.32e-05 | γδ-T.2 |
| ITGB2 | 2.7e-01 | 3.09e-01 | 8.62e-02 | 1.12e-04 | γδ-T.2 |
| 40787 | 3.06e-01 | 3.75e-01 | 3.05e-02 | 1.86e-04 | γδ-T.2 |
| PREX1 | 2.51e-01 | 2.62e-01 | 1.3e-01 | 3.71e-04 | γδ-T.2 |
| STARD3NL | 2.75e-01 | 3.98e-01 | 3.01e-02 | 3.92e-04 | γδ-T.2 |
| DENND1B | 3.41e-01 | 3.11e-01 | 1.77e-02 | 3.96e-04 | γδ-T.2 |

| | | | | | |
|---|---|---|---|---|---|
| ASXL2 | 2.76e-01 | 3.23e-01 | 2.6e-02 | 5.2e-04 | γδ-T.2 |
| GBE1 | 2.67e-01 | 7.04e-01 | 2.71e-03 | 5.71e-04 | γδ-T.2 |
| GLIPR2 | 2.64e-01 | 5.53e-01 | 6.8e-04 | 9.11e-04 | γδ-T.2 |
| 37681 | 3.6e-01 | 6.33e-01 | 3.49e-03 | 1.55e-03 | γδ-T.2 |
| ADCY3 | 2.54e-01 | 3.16e-01 | 6.09e-02 | 3.5e-03 | γδ-T.2 |
| SNTB1 | 2.61e-01 | 5.52e-01 | 1.91e-03 | 3.61e-03 | γδ-T.2 |
| ZNF767 | 2.54e-01 | 4.07e-01 | 4.71e-02 | 9.51e-03 | γδ-T.2 |
| A2M | 1.3e+00 | 1.57e+00 | 4.2e-17 | 5.7e-148 | γδ-T.3 |
| PLCB1 | 1.22e+00 | 1.58e+00 | 1.39e-13 | 8.61e-139 | γδ-T.3 |
| AV4S1 | 1.49e+00 | 1.88e+00 | 1.09e-21 | 1e-128 | γδ-T.3 |
| KLRB1 | 1.34e+00 | 1.89e+00 | 1.55e-17 | 2.46e-128 | γδ-T.3 |
| PZP | 1.13e+00 | 9.43e-01 | 1.36e-06 | 2.62e-114 | γδ-T.3 |
| S100B | 1.09e+00 | 1.08e+00 | 2.59e-13 | 1.26e-77 | γδ-T.3 |
| CEBPD | 8.84e-01 | 5.46e-01 | 1.49e-02 | 3.33e-71 | γδ-T.3 |
| RGS2 | 9.03e-01 | 4.24e-01 | 1.19e-01 | 1.74e-69 | γδ-T.3 |
| PDE4D | 8.17e-01 | 6.06e-01 | 8.48e-04 | 4.59e-58 | γδ-T.3 |
| KLRG1 | 7.68e-01 | 9.31e-01 | 6.04e-06 | 1.21e-56 | γδ-T.3 |
| ME1 | 6.58e-01 | 5.04e-01 | 6.24e-05 | 1.56e-56 | γδ-T.3 |
| CYTH3 | 8.5e-01 | 9.19e-01 | 4.78e-06 | 7.95e-56 | γδ-T.3 |
| RORA | 7.24e-01 | 1.11e+00 | 1.82e-06 | 5.56e-52 | γδ-T.3 |
| SLC4A10 | 6.32e-01 | 8.68e-01 | 1.73e-06 | 2.44e-47 | γδ-T.3 |
| IL12RB2 | 8.11e-01 | 1.14e+00 | 1.2e-07 | 1.86e-44 | γδ-T.3 |
| GZMK | 2.63e-01 | 9.87e-01 | 6.66e-08 | 3.15e-40 | γδ-T.3 |
| SLC7A5 | 6.11e-01 | 8.57e-01 | 2e-06 | 1.48e-38 | γδ-T.3 |
| RASGRF2 | 6.18e-01 | 3.76e-01 | 5.95e-02 | 1.77e-36 | γδ-T.3 |
| ABCB1 | 5.55e-01 | 7.6e-01 | 8.7e-03 | 4.62e-36 | γδ-T.3 |
| TARP | 4.96e-01 | 1.04e+00 | 8.4e-12 | 2.59e-34 | γδ-T.3 |
| STAT4 | 4.8e-01 | 1.05e+00 | 1.95e-07 | 7.92e-34 | γδ-T.3 |
| BC035094 | 5.86e-01 | 9.56e-01 | 4.13e-05 | 1.31e-33 | γδ-T.3 |
| PHACTR2 | 6.3e-01 | 1.18e+00 | 1.37e-09 | 2.21e-33 | γδ-T.3 |
| ARHGAP26 | 4.54e-01 | 1.07e+00 | 1.1e-12 | 2.89e-33 | γδ-T.3 |
| REL | 5.6e-01 | 3.3e-01 | 8.35e-02 | 8.42e-32 | γδ-T.3 |
| NFE2L2 | 5.71e-01 | 3.8e-01 | 1.02e-01 | 7.43e-31 | γδ-T.3 |
| SPOCK2 | 4.73e-01 | 6.11e-01 | 3.03e-03 | 1.34e-29 | γδ-T.3 |
| IFNGR1 | 5.69e-01 | 6.82e-01 | 8.89e-04 | 2.89e-29 | γδ-T.3 |
| CADM1 | 5.18e-01 | 2.9e-01 | 3.42e-02 | 6.33e-29 | γδ-T.3 |
| SATB1 | 5.99e-01 | 2.51e-01 | 3.44e-01 | 4.39e-28 | γδ-T.3 |
| NFKB1 | 6.07e-01 | 4.43e-01 | 3.09e-02 | 1.52e-26 | γδ-T.3 |

| | | | | | |
|---|---|---|---|---|---|
| PIK3AP1 | 5.44e-01 | 4.83e-01 | 2.56e-02 | 1.67e-24 | γδ-T.3 |
| DHRS3 | 5.58e-01 | 4.03e-01 | 2.55e-02 | 2.6e-24 | γδ-T.3 |
| DUSP2 | 4.48e-01 | 8.2e-01 | 1.25e-07 | 4.66e-24 | γδ-T.3 |
| NCR3 | 3.9e-01 | 4.9e-01 | 1.79e-02 | 3.21e-23 | γδ-T.3 |
| IL18RAP | 4.42e-01 | 7.49e-01 | 5.86e-09 | 4.39e-22 | γδ-T.3 |
| TTC39C | 3.95e-01 | 7.27e-01 | 1.91e-04 | 7.68e-22 | γδ-T.3 |
| LONRF3 | 4.76e-01 | 5.51e-01 | 1.76e-02 | 2.54e-21 | γδ-T.3 |
| MYBL1 | 5.03e-01 | 5.59e-01 | 6.22e-04 | 5.41e-21 | γδ-T.3 |
| PBX4 | 4.82e-01 | 1.11e+00 | 2.43e-10 | 9.5e-21 | γδ-T.3 |
| TMEM117 | 5.78e-01 | 6.6e-01 | 1.4e-02 | 5.63e-20 | γδ-T.3 |
| ERN1 | 4.34e-01 | 3.06e-01 | 3.48e-03 | 1.7e-19 | γδ-T.3 |
| NEO1 | 3.77e-01 | 6.54e-01 | 2.3e-05 | 3.43e-19 | γδ-T.3 |
| MAFF | 3.85e-01 | 5.13e-01 | 2.41e-02 | 1.87e-18 | γδ-T.3 |
| TANC2 | 4.57e-01 | 5.61e-01 | 5.34e-04 | 4.48e-18 | γδ-T.3 |
| FKBP11 | 4.14e-01 | 9.52e-01 | 1.01e-06 | 5.42e-18 | γδ-T.3 |
| DPP4 | 4.04e-01 | 6.72e-01 | 7.47e-06 | 8.08e-18 | γδ-T.3 |
| SVIL | 6.46e-01 | 6.66e-01 | 2.07e-02 | 9.86e-18 | γδ-T.3 |
| CRY1 | 4.59e-01 | 9.03e-01 | 4.17e-06 | 1.04e-16 | γδ-T.3 |
| LITAF | 4.79e-01 | 6.27e-01 | 2.8e-04 | 1.19e-16 | γδ-T.3 |
| PDE8A | 4.55e-01 | 3.79e-01 | 2.91e-02 | 1.27e-16 | γδ-T.3 |
| NFKBIA | 4.42e-01 | 8.21e-01 | 8.01e-05 | 1.52e-16 | γδ-T.3 |
| BC127952 | 3.92e-01 | 3.67e-01 | 1.15e-01 | 4.67e-16 | γδ-T.3 |
| MICAL2 | 4.49e-01 | 4.55e-01 | 6.21e-02 | 5.71e-16 | γδ-T.3 |
| TGFBR3 | 3.14e-01 | 7.16e-01 | 9.79e-04 | 9.59e-15 | γδ-T.3 |
| BTBD11 | 4.12e-01 | 5.01e-01 | 3.8e-02 | 2.63e-14 | γδ-T.3 |
| LBH | 3.54e-01 | 5.16e-01 | 5.87e-03 | 3.03e-14 | γδ-T.3 |
| GCHFR | 3.58e-01 | 9.74e-01 | 8.81e-09 | 4.9e-14 | γδ-T.3 |
| GALNT11 | 3.86e-01 | 8.85e-01 | 6.4e-05 | 4.76e-13 | γδ-T.3 |
| ATF7IP2 | 4.33e-01 | 3.09e-01 | 3.6e-02 | 1.77e-12 | γδ-T.3 |
| TC2N | 3.83e-01 | 6.36e-01 | 5.89e-03 | 2.56e-12 | γδ-T.3 |
| CD97 | 4.26e-01 | 8.72e-01 | 2.44e-05 | 2.85e-12 | γδ-T.3 |
| LOC285740 | 3.44e-01 | 5.02e-01 | 1.77e-03 | 3.74e-12 | γδ-T.3 |
| GPR65 | 3.7e-01 | 7.32e-01 | 9.37e-04 | 4.41e-12 | γδ-T.3 |
| EVA1C | 3.97e-01 | 8.61e-01 | 9.12e-07 | 4.67e-12 | γδ-T.3 |
| FBXO34 | 3.92e-01 | 6.54e-01 | 1.08e-03 | 7.29e-12 | γδ-T.3 |
| DUSP10 | 3.71e-01 | 5.45e-01 | 5.07e-02 | 9.4e-12 | γδ-T.3 |
| ZNF331 | 2.67e-01 | 2.63e-01 | 5.4e-02 | 1.75e-11 | γδ-T.3 |
| CDK17 | 2.97e-01 | 8.34e-01 | 1.62e-08 | 3.85e-11 | γδ-T.3 |

| | | | | | |
|---|---|---|---|---|---|
| ADRB2 | 2.73e-01 | 3.93e-01 | 1.44e-03 | 6.46e-11 | γδ-T.3 |
| ZFYVE9 | 2.83e-01 | 4.6e-01 | 3.35e-04 | 1.42e-10 | γδ-T.3 |
| SLC38A1 | 2.67e-01 | 4.9e-01 | 1.08e-02 | 2.89e-10 | γδ-T.3 |
| XBP1.1 | 2.95e-01 | 7.27e-01 | 3.26e-04 | 6.88e-10 | γδ-T.3 |
| SIK1 | 2.96e-01 | 2.51e-01 | 2.61e-01 | 9.25e-10 | γδ-T.3 |
| LYAR | 2.5e-01 | 1.15e+00 | 3.02e-06 | 9.75e-10 | γδ-T.3 |
| CEP112 | 3.75e-01 | 4.89e-01 | 1.05e-03 | 3.12e-09 | γδ-T.3 |
| LUZP1 | 3.12e-01 | 5.39e-01 | 4.46e-02 | 3.87e-09 | γδ-T.3 |
| SBF2 | 3.84e-01 | 3.84e-01 | 5.23e-02 | 5.94e-09 | γδ-T.3 |
| TESK2 | 3.5e-01 | 2.68e-01 | 1.45e-01 | 7.09e-09 | γδ-T.3 |
| ARNTL | 2.83e-01 | 5.81e-01 | 8.1e-03 | 7.78e-09 | γδ-T.3 |
| CERK | 2.97e-01 | 7.62e-01 | 3.93e-05 | 1.11e-08 | γδ-T.3 |
| FAM19A1 | 3.5e-01 | 2.79e-01 | 1.01e-01 | 1.9e-08 | γδ-T.3 |
| TM9SF1 | 2.72e-01 | 4.35e-01 | 3.33e-02 | 1.92e-08 | γδ-T.3 |
| EPHA4 | 2.77e-01 | 1.22e+00 | 1.34e-07 | 3.16e-08 | γδ-T.3 |
| B4GALT5 | 3.73e-01 | 1.11e+00 | 5.44e-06 | 4.6e-08 | γδ-T.3 |
| SND1 | 2.51e-01 | 3.55e-01 | 1.01e-01 | 8.54e-08 | γδ-T.3 |
| GZMM | 2.54e-01 | 4.41e-01 | 1.9e-02 | 1.14e-07 | γδ-T.3 |
| GYG1 | 2.78e-01 | 8.46e-01 | 2.43e-05 | 3.22e-07 | γδ-T.3 |
| SLA | 2.58e-01 | 4.39e-01 | 3.58e-02 | 3.28e-07 | γδ-T.3 |
| ZCCHC14 | 3.12e-01 | 3.87e-01 | 7.6e-02 | 4.63e-07 | γδ-T.3 |
| SESN1 | 2.73e-01 | 4.96e-01 | 3.79e-02 | 6.35e-07 | γδ-T.3 |
| MAP3K4 | 3.13e-01 | 4.33e-01 | 7.27e-03 | 6.93e-07 | γδ-T.3 |
| MPZL3 | 3.13e-01 | 4e-01 | 6.24e-03 | 3.94e-06 | γδ-T.3 |
| PDK3 | 2.66e-01 | 2.78e-01 | 2.37e-01 | 7.78e-06 | γδ-T.3 |
| DFNB31 | 2.64e-01 | 2.77e-01 | 3.11e-01 | 1.71e-05 | γδ-T.3 |
| GALNT10 | 2.62e-01 | 3.56e-01 | 1.85e-01 | 1.72e-05 | γδ-T.3 |
| VCL | 2.58e-01 | 3.3e-01 | 1.15e-01 | 5.41e-05 | γδ-T.3 |
| EPB41L2 | 3.03e-01 | 2.92e-01 | 3.94e-01 | 1.24e-04 | γδ-T.3 |
| TGFB1 | 2.61e-01 | 6.94e-01 | 1.48e-03 | 3.22e-04 | γδ-T.3 |
| FOS | 3.06e-01 | 3.56e-01 | 3.43e-01 | 3.22e-04 | γδ-T.3 |

**Table A.5:** List of differentially expressed genes between each of the γδ-T cell subtype and all other immune cell types in breast tumour sample BC1.

| gene | BC1 avg_logFC | p-value cluster | |
|---|---|---|---|
| FGFBP2 | 2.04e+00 | 9.66e-87 | γδ-T.1 |
| PRF1 | 1.72e+00 | 8.51e-53 | γδ-T.1 |
| RAP1GAP2 | 1.72e+00 | 1.2e-52 | γδ-T.1 |
| NKG7 | 1.5e+00 | 1.02e-43 | γδ-T.1 |
| GNLY | 1.46e+00 | 7.8e-37 | γδ-T.1 |
| SPON2 | 1.07e+00 | 2.25e-33 | γδ-T.1 |
| S1PR5 | 9.44e-01 | 8.76e-30 | γδ-T.1 |
| PLAC8 | 1.14e+00 | 4.22e-27 | γδ-T.1 |
| GZMB | 1.17e+00 | 1.67e-25 | γδ-T.1 |
| GZMH | 1.06e+00 | 6.46e-24 | γδ-T.1 |
| LYN | 1.17e+00 | 1.29e-19 | γδ-T.1 |
| UBE2F | 1.11e+00 | 1.7e-19 | γδ-T.1 |
| KLRF1 | 1.13e+00 | 3.73e-18 | γδ-T.1 |
| PDGFD | 1.39e+00 | 9.68e-18 | γδ-T.1 |
| KLRD1 | 8.65e-01 | 2.61e-17 | γδ-T.1 |
| BNC2 | 1.1e+00 | 5.8e-17 | γδ-T.1 |
| FAM53B | 9.37e-01 | 1.13e-16 | γδ-T.1 |
| CD247 | 8.49e-01 | 1.49e-16 | γδ-T.1 |
| TGFBR3 | 1e+00 | 1.44e-15 | γδ-T.1 |
| ZEB2 | 7.37e-01 | 4.42e-15 | γδ-T.1 |
| TMCC3 | 9.45e-01 | 8.42e-15 | γδ-T.1 |
| RAP2B | 8.22e-01 | 2.68e-14 | γδ-T.1 |
| TTC38 | 7.48e-01 | 6.37e-14 | γδ-T.1 |
| PTPN12 | 9.16e-01 | 6.46e-14 | γδ-T.1 |
| TFDP2 | 8.86e-01 | 2.84e-13 | γδ-T.1 |
| KLHDC4 | 7.14e-01 | 1.31e-12 | γδ-T.1 |
| BCR | 7.11e-01 | 2.26e-12 | γδ-T.1 |
| GK5 | 8.35e-01 | 5.38e-12 | γδ-T.1 |
| AUTS2 | 8.48e-01 | 2.5e-11 | γδ-T.1 |
| DGKD | 7.75e-01 | 5.93e-11 | γδ-T.1 |
| NCAM1 | 8.34e-01 | 1.47e-10 | γδ-T.1 |
| SGCD | 7.69e-01 | 1.95e-10 | γδ-T.1 |
| TYROBP | 7.1e-01 | 9.85e-10 | γδ-T.1 |
| DOCK5 | 1.02e+00 | 1.01e-09 | γδ-T.1 |

| | | | |
|---|---|---|---|
| RASA3 | 7.52e-01 | 1.82e-09 | γδ-T.1 |
| PTPN7 | 7.59e-01 | 2.11e-09 | γδ-T.1 |
| MYO1F | 7.32e-01 | 3.06e-09 | γδ-T.1 |
| B3GNT7 | 7.43e-01 | 3.35e-09 | γδ-T.1 |
| ITGAL | 7.43e-01 | 3.97e-09 | γδ-T.1 |
| MTSS1 | 7.03e-01 | 7.21e-09 | γδ-T.1 |
| HAVCR2 | 7.56e-01 | 9.11e-09 | γδ-T.1 |
| SSBP3 | 7.13e-01 | 1.47e-08 | γδ-T.1 |
| YES1 | 9.44e-01 | 3.04e-08 | γδ-T.1 |
| CADM1 | 7.55e-01 | 3.63e-08 | γδ-T.1 |
| FOXK2 | 7.68e-01 | 3.97e-08 | γδ-T.1 |
| LINGO2 | 1.43e+00 | 8.16e-08 | γδ-T.1 |
| IGF1R | 8.57e-01 | 1.33e-05 | γδ-T.1 |
| GNLY | 3.16e+00 | 3.98e-246 | γδ-T.2 |
| ATP8B4 | 2.02e+00 | 1.71e-177 | γδ-T.2 |
| NCAM1 | 1.97e+00 | 1.48e-163 | γδ-T.2 |
| TYROBP | 1.41e+00 | 1.35e-130 | γδ-T.2 |
| PLCG2 | 2.02e+00 | 6.43e-124 | γδ-T.2 |
| KLRD1 | 1.45e+00 | 1.4e-101 | γδ-T.2 |
| NKG7 | 1.37e+00 | 1.19e-89 | γδ-T.2 |
| CTSW | 1.13e+00 | 1.85e-75 | γδ-T.2 |
| RIN3 | 1.35e+00 | 4.1e-73 | γδ-T.2 |
| SH2D1B | 8.52e-01 | 4.95e-73 | γδ-T.2 |
| KLRC1 | 1.2e+00 | 8e-73 | γδ-T.2 |
| GZMB | 1.68e+00 | 5.05e-70 | γδ-T.2 |
| NCR1 | 7.35e-01 | 1.75e-68 | γδ-T.2 |
| FCER1G | 9.3e-01 | 3.56e-68 | γδ-T.2 |
| CLNK | 1e+00 | 6.41e-62 | γδ-T.2 |
| MCTP2 | 1.15e+00 | 1.89e-61 | γδ-T.2 |
| LINGO2 | 1.88e+00 | 2.57e-60 | γδ-T.2 |
| HAVCR2 | 1.25e+00 | 5.47e-60 | γδ-T.2 |
| NCALD | 1.16e+00 | 3.73e-58 | γδ-T.2 |
| CD7 | 1.04e+00 | 2.2e-57 | γδ-T.2 |
| KLRF1 | 9.29e-01 | 1.76e-55 | γδ-T.2 |
| B3GNT7 | 7.5e-01 | 2.24e-53 | γδ-T.2 |
| CNOT2 | 1.15e+00 | 6.03e-51 | γδ-T.2 |
| CD63 | 9.39e-01 | 1.12e-50 | γδ-T.2 |
| KLRK1 | 8.22e-01 | 1.16e-48 | γδ-T.2 |

| | | | |
|---|---|---|---|
| DOCK5 | 1.1e+00 | 4.37e-48 | γδ-T.2 |
| PRF1 | 1.06e+00 | 7.6e-48 | γδ-T.2 |
| GOLIM4 | 7.73e-01 | 1.88e-45 | γδ-T.2 |
| GAS7 | 1.1e+00 | 6.99e-45 | γδ-T.2 |
| APBA2 | 1.03e+00 | 1.39e-42 | γδ-T.2 |
| PRKX | 8.38e-01 | 2.56e-42 | γδ-T.2 |
| VAV3 | 1.04e+00 | 2.95e-41 | γδ-T.2 |
| LYN | 8.57e-01 | 3.52e-41 | γδ-T.2 |
| HSH2D | 9.75e-01 | 9.18e-41 | γδ-T.2 |
| PDGFD | 1.06e+00 | 7.29e-40 | γδ-T.2 |
| ABCB1 | 9.09e-01 | 2.6e-39 | γδ-T.2 |
| SAMD3 | 8.4e-01 | 5.55e-39 | γδ-T.2 |
| TXK | 8.87e-01 | 1.13e-38 | γδ-T.2 |
| HOPX | 9.69e-01 | 1.75e-38 | γδ-T.2 |
| GEM | 1.12e+00 | 7.84e-38 | γδ-T.2 |
| GALNT2 | 9.14e-01 | 8.88e-38 | γδ-T.2 |
| PPP1R9A | 7.12e-01 | 2.47e-37 | γδ-T.2 |
| ADAM28 | 8.03e-01 | 3.11e-37 | γδ-T.2 |
| SRGAP3 | 8.84e-01 | 5.02e-34 | γδ-T.2 |
| ABHD2 | 8.46e-01 | 8.79e-34 | γδ-T.2 |
| AHI1 | 8.92e-01 | 4.68e-33 | γδ-T.2 |
| ITGB1 | 7.9e-01 | 1.12e-32 | γδ-T.2 |
| PDE7B | 1.14e+00 | 2.87e-32 | γδ-T.2 |
| ZNF683 | 7.92e-01 | 1.31e-30 | γδ-T.2 |
| PIK3AP1 | 8.8e-01 | 1.75e-30 | γδ-T.2 |
| LARGE | 9.12e-01 | 4.08e-30 | γδ-T.2 |
| LDB2 | 7.59e-01 | 2.5e-28 | γδ-T.2 |
| IFITM2 | 7.06e-01 | 6.28e-28 | γδ-T.2 |
| ITGA1 | 7.6e-01 | 4.48e-27 | γδ-T.2 |
| LOC374443 | 7.05e-01 | 2.23e-25 | γδ-T.2 |
| L3MBTL4 | 8e-01 | 1.71e-24 | γδ-T.2 |
| AGPAT4 | 7.38e-01 | 4.34e-23 | γδ-T.2 |
| KCNQ5 | 9.67e-01 | 7.78e-22 | γδ-T.2 |
| CMC1 | 9.59e-01 | 3.89e-20 | γδ-T.2 |
| DAPK2 | 7.77e-01 | 2.25e-18 | γδ-T.2 |
| ABTB2 | 8.52e-01 | 1.02e-10 | γδ-T.2 |
| A2M | 1.2e+00 | 1.28e-84 | γδ-T.3 |
| ZBTB16 | 1.09e+00 | 3.84e-77 | γδ-T.3 |

| | | | |
|---|---|---|---|
| PLCB1 | 1.09e+00 | 5.28e-76 | γδ-T.3 |
| PZP | 1.05e+00 | 7.8e-64 | γδ-T.3 |
| KLRB1 | 9.95e-01 | 4.99e-54 | γδ-T.3 |
| IL12RB2 | 1.07e+00 | 3.59e-52 | γδ-T.3 |
| CEBPD | 8.32e-01 | 7.24e-41 | γδ-T.3 |
| KLRD1 | 7.88e-01 | 1.26e-40 | γδ-T.3 |
| RGS2 | 8.16e-01 | 4.34e-36 | γδ-T.3 |
| CYTH3 | 8.78e-01 | 5.25e-36 | γδ-T.3 |
| S100B | 7.63e-01 | 4.4e-27 | γδ-T.3 |

**Table A.6:** List of GO Biological Process and KEGG pathway terms significantly enriched in the γδ-T cell clusters in PBMC.

| Term | Fold Enrichment | Bonferroni | cluster |
|---|---|---|---|
| GO:0006614 SRP-dependent cotranslational protein targeting to membrane | 6.26e+01 | 1.08e-29 | δ1.1 |
| GO:0006613 cotranslational protein targeting to membrane | 5.83e+01 | 5.55e-29 | δ1.1 |
| GO:0045047 protein targeting to ER | 5.77e+01 | 6.94e-29 | δ1.1 |
| GO:0072599 establishment of protein localization to endoplasmic reticulum | 5.56e+01 | 1.65e-28 | δ1.1 |
| GO:0000184 nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 4.88e+01 | 3.21e-27 | δ1.1 |
| GO:0070972 protein localization to endoplasmic reticulum | 4.69e+01 | 7.9e-27 | δ1.1 |
| GO:0019083 viral transcription | 3.47e+01 | 2.95e-25 | δ1.1 |
| GO:0019080 viral gene expression | 3.27e+01 | 1.14e-24 | δ1.1 |
| GO:0044033 multi-organism metabolic process | 2.93e+01 | 1.35e-23 | δ1.1 |
| GO:0006413 translational initiation | 3.18e+01 | 4.38e-23 | δ1.1 |
| GO:0006612 protein targeting to membrane | 3.16e+01 | 4.92e-23 | δ1.1 |
| hsa03010:Ribosome | 2.59e+01 | 1.27e-23 | δ1.1 |
| GO:0000956 nuclear-transcribed mRNA catabolic process | 2.92e+01 | 2.84e-22 | δ1.1 |
| GO:0006402 mRNA catabolic process | 2.72e+01 | 1.29e-21 | δ1.1 |
| GO:0006401 RNA catabolic process | 2.41e+01 | 1.65e-20 | δ1.1 |
| GO:0006364 rRNA processing | 2.23e+01 | 8.54e-20 | δ1.1 |
| GO:0016072 rRNA metabolic process | 2.18e+01 | 1.47e-19 | δ1.1 |
| GO:0042254 ribosome biogenesis | 1.79e+01 | 8.59e-18 | δ1.1 |
| GO:0019058 viral life cycle | 1.44e+01 | 9.99e-18 | δ1.1 |
| GO:0072657 protein localization to membrane | 1.34e+01 | 5.05e-17 | δ1.1 |
| GO:0090150 establishment of protein localization to membrane | 1.63e+01 | 6.05e-17 | δ1.1 |
| GO:0006412 translation | 1.11e+01 | 6.76e-17 | δ1.1 |

| | | | |
|---|---|---|---|
| GO:0034655 nucleobase-containing compound catabolic process | 1.58e+01 | 1.18e-16 | δ1.1 |
| GO:0043043 peptide biosynthetic process | 1.06e+01 | 1.76e-16 | δ1.1 |
| GO:0034470 ncRNA processing | 1.46e+01 | 5.92e-16 | δ1.1 |
| GO:0046700 heterocycle catabolic process | 1.46e+01 | 5.92e-16 | δ1.1 |
| GO:0044270 cellular nitrogen compound catabolic process | 1.44e+01 | 8.37e-16 | δ1.1 |
| GO:0019439 aromatic compound catabolic process | 1.42e+01 | 1.07e-15 | δ1.1 |
| GO:0043604 amide biosynthetic process | 9.64e+00 | 1.85e-15 | δ1.1 |
| GO:1901361 organic cyclic compound catabolic process | 1.35e+01 | 3.01e-15 | δ1.1 |
| GO:0044802 single-organism membrane organization | 8.51e+00 | 6.19e-15 | δ1.1 |
| GO:1902582 single-organism intracellular transport | 9.82e+00 | 7.76e-15 | δ1.1 |
| GO:0022613 ribonucleoprotein complex biogenesis | 1.25e+01 | 1.41e-14 | δ1.1 |
| GO:0006518 peptide metabolic process | 8.68e+00 | 2.26e-14 | δ1.1 |
| GO:0072594 establishment of protein localization to organelle | 9.84e+00 | 4.71e-14 | δ1.1 |
| GO:0016032 viral process | 7.63e+00 | 8.89e-14 | δ1.1 |
| GO:0044764 multi-organism cellular process | 7.57e+00 | 1.98e-13 | δ1.1 |
| GO:0006605 protein targeting | 9.34e+00 | 1.98e-13 | δ1.1 |
| GO:0044403 symbiosis, encompassing mutualism through parasitism | 7.38e+00 | 1.98e-13 | δ1.1 |
| GO:0044419 interspecies interaction between organisms | 7.38e+00 | 1.98e-13 | δ1.1 |
| GO:0061024 membrane organization | 7.18e+00 | 3.96e-13 | δ1.1 |
| GO:0034660 ncRNA metabolic process | 1.05e+01 | 3.96e-13 | δ1.1 |
| GO:0043603 cellular amide metabolic process | 7.14e+00 | 2.38e-12 | δ1.1 |
| GO:0016071 mRNA metabolic process | 8.94e+00 | 1.27e-11 | δ1.1 |
| GO:1902580 single-organism cellular localization | 6.44e+00 | 2.5e-11 | δ1.1 |
| GO:0033365 protein localization to organelle | 7.26e+00 | 3.51e-11 | δ1.1 |

| | | | |
|---|---|---|---|
| GO:1901566 organonitrogen compound biosynthetic process | 5.36e+00 | 4.37e-10 | δ1.1 |
| GO:0006886 intracellular protein transport | 6.42e+00 | 4.72e-10 | δ1.1 |
| GO:0006396 RNA processing | 6.6e+00 | 4.59e-09 | δ1.1 |
| GO:0044265 cellular macromolecule catabolic process | 6.15e+00 | 1.74e-08 | δ1.1 |
| GO:0015031 protein transport | 4.11e+00 | 6.5e-08 | δ1.1 |
| GO:0045184 establishment of protein localization | 3.92e+00 | 7e-08 | δ1.1 |
| GO:0046907 intracellular transport | 4.28e+00 | 6.36e-07 | δ1.1 |
| GO:0034613 cellular protein localization | 4.26e+00 | 6.87e-07 | δ1.1 |
| GO:0070727 cellular macromolecule localization | 4.23e+00 | 8.11e-07 | δ1.1 |
| GO:0009057 macromolecule catabolic process | 5e+00 | 8.28e-07 | δ1.1 |
| GO:0008104 protein localization | 3.34e+00 | 1.16e-06 | δ1.1 |
| GO:0051649 establishment of localization in cell | 3.68e+00 | 2.16e-06 | δ1.1 |
| GO:0033036 macromolecule localization | 2.91e+00 | 3.21e-05 | δ1.1 |
| GO:0044085 cellular component biogenesis | 2.81e+00 | 6.99e-05 | δ1.1 |
| GO:0051641 cellular localization | 3.01e+00 | 7.08e-05 | δ1.1 |
| GO:0044248 cellular catabolic process | 3.72e+00 | 1.62e-04 | δ1.1 |
| GO:0032774 RNA biosynthetic process | 2.24e+00 | 3.44e-03 | δ1.1 |
| GO:0010467 gene expression | 1.9e+00 | 1.15e-02 | δ1.1 |
| GO:0034654 nucleobase-containing compound biosynthetic process | 2.07e+00 | 1.19e-02 | δ1.1 |
| GO:0018130 heterocycle biosynthetic process | 2.04e+00 | 1.58e-02 | δ1.1 |
| GO:0019438 aromatic compound biosynthetic process | 2.04e+00 | 1.68e-02 | δ1.1 |
| GO:0034645 cellular macromolecule biosynthetic process | 1.91e+00 | 2.91e-02 | δ1.1 |
| GO:0048534 hematopoietic or lymphoid organ development | 4.56e+00 | 3.09e-02 | δ1.1 |
| GO:0002520 immune system development | 4.32e+00 | 5.2e-02 | δ1.1 |
| GO:0002181 cytoplasmic translation | 2.58e+01 | 6.74e-02 | δ1.1 |

| | | | |
|---|---|---|---|
| GO:0016337 single organismal cell-cell adhesion | 4.52e+00 | 8.24e-02 | δ1.1 |
| GO:0098609 cell-cell adhesion | 3.52e+00 | 8.7e-02 | δ1.1 |
| GO:0030097 hemopoiesis | 4.43e+00 | 9.71e-02 | δ1.1 |
| GO:0016337 single organismal cell-cell adhesion | 6.35e+00 | 7.03e-04 | δ1.2 |
| GO:0001775 cell activation | 5.45e+00 | 1.12e-03 | δ1.2 |
| GO:0098602 single organism cell adhesion | 5.91e+00 | 1.51e-03 | δ1.2 |
| GO:0030217 T cell differentiation | 1.39e+01 | 2.22e-03 | δ1.2 |
| GO:0030098 lymphocyte differentiation | 1.03e+01 | 3.1e-03 | δ1.2 |
| GO:0042110 T cell activation | 7.94e+00 | 4.97e-03 | δ1.2 |
| GO:0070489 T cell aggregation | 7.94e+00 | 4.97e-03 | δ1.2 |
| GO:0071593 lymphocyte aggregation | 7.93e+00 | 5.07e-03 | δ1.2 |
| GO:0070486 leukocyte aggregation | 7.8e+00 | 5.75e-03 | δ1.2 |
| GO:0048534 hematopoietic or lymphoid organ development | 5.46e+00 | 1.06e-02 | δ1.2 |
| GO:0007159 leukocyte cell-cell adhesion | 7.21e+00 | 1.08e-02 | δ1.2 |
| GO:0046649 lymphocyte activation | 6.06e+00 | 1.32e-02 | δ1.2 |
| GO:0002520 immune system development | 5.18e+00 | 1.77e-02 | δ1.2 |
| GO:0030097 hemopoiesis | 5.27e+00 | 4.34e-02 | δ1.2 |
| GO:0045321 leukocyte activation | 5.22e+00 | 4.75e-02 | δ1.2 |
| GO:0002521 leukocyte differentiation | 6.73e+00 | 6.6e-02 | δ1.2 |
| GO:0042127 regulation of cell proliferation | 3.42e+00 | 7.73e-02 | δ1.2 |
| GO:0098609 cell-cell adhesion | 3.96e+00 | 8.61e-02 | δ1.2 |
| GO:0006955 immune response | 4.6e+00 | 1.76e-14 | δ2.1 |
| GO:0002252 immune effector process | 6.41e+00 | 2.17e-11 | δ2.1 |
| GO:0048584 positive regulation of response to stimulus | 3.26e+00 | 3.27e-08 | δ2.1 |
| GO:0050776 regulation of immune response | 4.84e+00 | 1.33e-07 | δ2.1 |
| GO:0006909 phagocytosis | 9.81e+00 | 1.76e-07 | δ2.1 |
| GO:0001775 cell activation | 4.67e+00 | 7.74e-07 | δ2.1 |
| GO:0006897 endocytosis | 5.37e+00 | 3.31e-06 | δ2.1 |
| GO:0002682 regulation of immune system process | 3.58e+00 | 7.03e-06 | δ2.1 |
| GO:0006928 movement of cell or subcellular component | 3.12e+00 | 9.11e-06 | δ2.1 |
| GO:0016192 vesicle-mediated transport | 3.41e+00 | 9.52e-06 | δ2.1 |

| | | | |
|---|---|---|---|
| GO:0006952 defense response | 3.36e+00 | 1.37e-05 | δ2.1 |
| GO:0007155 cell adhesion | 3.1e+00 | 4.24e-05 | δ2.1 |
| GO:0022610 biological adhesion | 3.09e+00 | 4.59e-05 | δ2.1 |
| GO:0007166 cell surface receptor signaling pathway | 2.48e+00 | 9.21e-05 | δ2.1 |
| GO:0016477 cell migration | 3.59e+00 | 1.36e-04 | δ2.1 |
| GO:0002433 immune response-regulating cell surface receptor signaling pathway involved in phagocytosis | 1.34e+01 | 1.45e-04 | δ2.1 |
| GO:0038096 Fc-gamma receptor signaling pathway involved in phagocytosis | 1.34e+01 | 1.45e-04 | δ2.1 |
| GO:0002431 Fc receptor mediated stimulatory signaling pathway | 1.31e+01 | 1.77e-04 | δ2.1 |
| GO:0038094 Fc-gamma receptor signaling pathway | 1.3e+01 | 1.9e-04 | δ2.1 |
| GO:0050778 positive regulation of immune response | 4.72e+00 | 1.94e-04 | δ2.1 |
| GO:0040011 locomotion | 3.11e+00 | 2.85e-04 | δ2.1 |
| GO:0051674 localization of cell | 3.32e+00 | 2.9e-04 | δ2.1 |
| GO:0048870 cell motility | 3.32e+00 | 2.9e-04 | δ2.1 |
| GO:0002684 positive regulation of immune system process | 3.89e+00 | 3.14e-04 | δ2.1 |
| GO:0002757 immune response-activating signal transduction | 5.46e+00 | 4.91e-04 | δ2.1 |
| GO:0007229 integrin-mediated signaling pathway | 1.44e+01 | 5.07e-04 | δ2.1 |
| GO:0002764 immune response-regulating signaling pathway | 5.12e+00 | 1.11e-03 | δ2.1 |
| GO:0098602 single organism cell adhesion | 4.14e+00 | 1.33e-03 | δ2.1 |
| GO:0002253 activation of immune response | 4.94e+00 | 1.74e-03 | δ2.1 |
| GO:0016337 single organismal cell-cell adhesion | 4.22e+00 | 2.36e-03 | δ2.1 |
| GO:0045321 leukocyte activation | 4.09e+00 | 3.56e-03 | δ2.1 |
| GO:0002429 immune response-activating cell surface receptor signaling pathway | 6.02e+00 | 3.94e-03 | δ2.1 |
| hsa04670:Leukocyte transendothelial migration | 1.04e+01 | 2.51e-04 | δ2.1 |

| | | | |
|---|---|---|---|
| hsa04650:Natural killer cell mediated cytotoxicity | 9.76e+00 | 3.92e-04 | δ2.1 |
| GO:0002768 immune response-regulating cell surface receptor signaling pathway | 5.54e+00 | 9.34e-03 | δ2.1 |
| GO:0043207 response to external biotic stimulus | 3.69e+00 | 1.44e-02 | δ2.1 |
| GO:0051707 response to other organism | 3.69e+00 | 1.44e-02 | δ2.1 |
| GO:0098609 cell-cell adhesion | 3.07e+00 | 2.71e-02 | δ2.1 |
| GO:0009607 response to biotic stimulus | 3.5e+00 | 2.83e-02 | δ2.1 |
| GO:0050900 leukocyte migration | 5.42e+00 | 3.15e-02 | δ2.1 |
| GO:0038093 Fc receptor signaling pathway | 6.99e+00 | 3.24e-02 | δ2.1 |
| hsa04810:Regulation of actin cytoskeleton | 6.3e+00 | 2.9e-03 | δ2.1 |
| GO:0046649 lymphocyte activation | 3.96e+00 | 5.44e-02 | δ2.1 |
| GO:0007159 leukocyte cell-cell adhesion | 4.5e+00 | 7.25e-02 | δ2.1 |
| GO:0006968 cellular defense response | 1.64e+01 | 7.85e-02 | δ2.1 |
| GO:0019835 cytolysis | 2.74e+01 | 7.87e-02 | δ2.1 |
| GO:0009605 response to external stimulus | 2.29e+00 | 9.28e-02 | δ2.1 |
| GO:0002576 platelet degranulation | 1.11e+01 | 9.72e-02 | δ2.1 |
| hsa05131:Shigellosis | 1.24e+01 | 1.5e-02 | δ2.1 |
| hsa04015:Rap1 signaling pathway | 5.67e+00 | 2.02e-02 | δ2.1 |
| hsa05140:Leishmaniasis | 1.12e+01 | 2.45e-02 | δ2.1 |
| hsa04142:Lysosome | 7.65e+00 | 3.63e-02 | δ2.1 |
| hsa05132:Salmonella infection | 9.56e+00 | 5.03e-02 | δ2.1 |
| GO:0007166 cell surface receptor signaling pathway | 4.68e+00 | 1.1e-03 | δ2.2 |
| GO:0006952 defense response | 6.86e+00 | 1.1e-03 | δ2.2 |
| GO:0019221 cytokine-mediated signaling pathway | 1.31e+01 | 4.38e-03 | δ2.2 |
| GO:0002682 regulation of immune system process | 6.8e+00 | 6.99e-03 | δ2.2 |
| GO:0071345 cellular response to cytokine stimulus | 1.03e+01 | 1.74e-02 | δ2.2 |
| GO:0006968 cellular defense response | 6.71e+01 | 1.84e-02 | δ2.2 |
| GO:1903039 positive regulation of leukocyte cell-cell adhesion | 2.39e+01 | 2.95e-02 | δ2.2 |
| GO:0034097 response to cytokine | 9.08e+00 | 3.56e-02 | δ2.2 |
| GO:0022409 positive regulation of cell-cell adhesion | 2.1e+01 | 4.82e-02 | δ2.2 |

| | | | |
|---|---|---|---|
| GO:0050776 regulation of immune response | 7.99e+00 | 7.17e-02 | δ2.2 |
| hsa04060:Cytokine-cytokine receptor interaction | 1.42e+01 | 4.81e-03 | δ2.2 |
| GO:0006614 SRP-dependent cotranslational protein targeting to membrane | 5.61e+01 | 2.26e-12 | δ2.3 |
| GO:0006613 cotranslational protein targeting to membrane | 5.23e+01 | 4.52e-12 | δ2.3 |
| GO:0045047 protein targeting to ER | 5.18e+01 | 5.12e-12 | δ2.3 |
| GO:0072599 establishment of protein localization to endoplasmic reticulum | 4.99e+01 | 7.62e-12 | δ2.3 |
| GO:0000184 nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 4.38e+01 | 2.87e-11 | δ2.3 |
| GO:0070972 protein localization to endoplasmic reticulum | 4.21e+01 | 4.31e-11 | δ2.3 |
| GO:0019083 viral transcription | 2.98e+01 | 1.43e-09 | δ2.3 |
| GO:0006413 translational initiation | 2.85e+01 | 2.21e-09 | δ2.3 |
| GO:0006612 protein targeting to membrane | 2.84e+01 | 2.33e-09 | δ2.3 |
| GO:0019080 viral gene expression | 2.81e+01 | 2.59e-09 | δ2.3 |
| hsa03010:Ribosome | 2.42e+01 | 6.22e-11 | δ2.3 |
| GO:0000956 nuclear-transcribed mRNA catabolic process | 2.61e+01 | 5.23e-09 | δ2.3 |
| GO:0044033 multi-organism metabolic process | 2.52e+01 | 7.64e-09 | δ2.3 |
| GO:0006402 mRNA catabolic process | 2.44e+01 | 1.05e-08 | δ2.3 |
| GO:0006401 RNA catabolic process | 2.16e+01 | 3.43e-08 | δ2.3 |
| GO:0006364 rRNA processing | 2e+01 | 7.34e-08 | δ2.3 |
| GO:0016072 rRNA metabolic process | 1.95e+01 | 9.44e-08 | δ2.3 |
| GO:0042254 ribosome biogenesis | 1.61e+01 | 6.26e-07 | δ2.3 |
| GO:0019058 viral life cycle | 1.29e+01 | 6.79e-07 | δ2.3 |
| GO:0090150 establishment of protein localization to membrane | 1.46e+01 | 1.55e-06 | δ2.3 |
| GO:0034655 nucleobase-containing compound catabolic process | 1.42e+01 | 2.12e-06 | δ2.3 |
| GO:0046700 heterocycle catabolic process | 1.31e+01 | 4.5e-06 | δ2.3 |

| | | | |
|---|---|---|---|
| GO:0034470 ncRNA processing | 1.31e+01 | 4.5e-06 | δ2.3 |
| GO:0044270 cellular nitrogen compound catabolic process | 1.29e+01 | 5.29e-06 | δ2.3 |
| GO:0019439 aromatic compound catabolic process | 1.27e+01 | 5.93e-06 | δ2.3 |
| GO:1901361 organic cyclic compound catabolic process | 1.21e+01 | 9.62e-06 | δ2.3 |
| GO:0022613 ribonucleoprotein complex biogenesis | 1.12e+01 | 1.98e-05 | δ2.3 |
| GO:0072657 protein localization to membrane | 1.1e+01 | 2.36e-05 | δ2.3 |
| GO:0006412 translation | 9.17e+00 | 2.42e-05 | δ2.3 |
| GO:0006518 peptide metabolic process | 7.78e+00 | 2.67e-05 | δ2.3 |
| GO:0043043 peptide biosynthetic process | 8.81e+00 | 3.64e-05 | δ2.3 |
| GO:0016071 mRNA metabolic process | 8.74e+00 | 3.93e-05 | δ2.3 |
| GO:0043604 amide biosynthetic process | 7.98e+00 | 9.94e-05 | δ2.3 |
| GO:0034660 ncRNA metabolic process | 9.45e+00 | 9.96e-05 | δ2.3 |
| GO:0016032 viral process | 6.59e+00 | 1.66e-04 | δ2.3 |
| GO:0044764 multi-organism cellular process | 6.54e+00 | 1.8e-04 | δ2.3 |
| GO:0043603 cellular amide metabolic process | 6.4e+00 | 2.27e-04 | δ2.3 |
| GO:0044403 symbiosis, encompassing mutualism through parasitism | 6.38e+00 | 2.37e-04 | δ2.3 |
| GO:0044419 interspecies interaction between organisms | 6.38e+00 | 2.37e-04 | δ2.3 |
| GO:0072594 establishment of protein localization to organelle | 8.09e+00 | 4.22e-04 | δ2.3 |
| GO:1902582 single-organism intracellular transport | 7.75e+00 | 6.24e-04 | δ2.3 |
| GO:0006605 protein targeting | 7.68e+00 | 6.75e-04 | δ2.3 |
| GO:0006396 RNA processing | 6.46e+00 | 8.36e-04 | δ2.3 |
| GO:0044265 cellular macromolecule catabolic process | 6.02e+00 | 1.67e-03 | δ2.3 |
| GO:0044802 single-organism membrane organization | 6.22e+00 | 4.56e-03 | δ2.3 |
| GO:0033365 protein localization to organelle | 5.97e+00 | 6.54e-03 | δ2.3 |

| | | | |
|---|---|---|---|
| GO:0009057 macromolecule catabolic process | 4.89e+00 | 1.23e-02 | δ2.3 |
| GO:0006886 intracellular protein transport | 5.28e+00 | 1.92e-02 | δ2.3 |
| GO:0061024 membrane organization | 5.25e+00 | 2.03e-02 | δ2.3 |
| GO:0070489 T cell aggregation | 8.78e+00 | 2.34e-02 | δ2.3 |
| GO:0042110 T cell activation | 8.78e+00 | 2.34e-02 | δ2.3 |
| GO:0071593 lymphocyte aggregation | 8.76e+00 | 2.37e-02 | δ2.3 |
| GO:0070486 leukocyte aggregation | 8.63e+00 | 2.62e-02 | δ2.3 |
| GO:0001913 T cell mediated cytotoxicity | 6.53e+01 | 2.93e-02 | δ2.3 |
| GO:0046649 lymphocyte activation | 6.85e+00 | 2.95e-02 | δ2.3 |
| GO:1902580 single-organism cellular localization | 4.89e+00 | 3.7e-02 | δ2.3 |
| GO:0007159 leukocyte cell-cell adhesion | 7.98e+00 | 4.28e-02 | δ2.3 |
| GO:1901566 organonitrogen compound biosynthetic process | 4.27e+00 | 4.32e-02 | δ2.3 |
| GO:0002682 regulation of immune system process | 4.26e+00 | 4.4e-02 | δ2.3 |
| GO:0098609 cell-cell adhesion | 4.63e+00 | 5.85e-02 | δ2.3 |
| GO:0045321 leukocyte activation | 5.9e+00 | 8.3e-02 | δ2.3 |

**Table A.7:** List of GO Biological Process and KEGG pathway terms differentially enriched between δ1.1 and δ1.2 γδ-T cell subtypes in PBMC.

| Term | Fold Enrichment | Bonferroni | cluster |
|---|---|---|---|
| GO:0006614 SRP-dependent cotranslational protein targeting to membrane | 1.39e+02 | 8.83e-97 | δ1.1 |
| GO:0006613 cotranslational protein targeting to membrane | 1.29e+02 | 7.51e-95 | δ1.1 |
| GO:0045047 protein targeting to ER | 1.28e+02 | 1.37e-94 | δ1.1 |
| GO:0072599 establishment of protein localization to endoplasmic reticulum | 1.23e+02 | 1.4e-93 | δ1.1 |
| GO:0000184 nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 1.08e+02 | 3.42e-90 | δ1.1 |
| GO:0070972 protein localization to endoplasmic reticulum | 1.04e+02 | 3.52e-89 | δ1.1 |
| GO:0019083 viral transcription | 7.36e+01 | 9.29e-81 | δ1.1 |
| GO:0006413 translational initiation | 7.05e+01 | 9.57e-80 | δ1.1 |
| GO:0006612 protein targeting to membrane | 7.01e+01 | 1.27e-79 | δ1.1 |
| GO:0019080 viral gene expression | 6.94e+01 | 2.23e-79 | δ1.1 |
| GO:0000956 nuclear-transcribed mRNA catabolic process | 6.47e+01 | 9.5e-78 | δ1.1 |
| GO:0044033 multi-organism metabolic process | 6.22e+01 | 7.13e-77 | δ1.1 |
| GO:0006402 mRNA catabolic process | 6.03e+01 | 3.89e-76 | δ1.1 |
| GO:0006401 RNA catabolic process | 5.35e+01 | 1.96e-73 | δ1.1 |
| hsa03010:Ribosome | 4.34e+01 | 3.45e-74 | δ1.1 |
| GO:0006364 rRNA processing | 4.95e+01 | 1.05e-71 | δ1.1 |
| GO:0016072 rRNA metabolic process | 4.83e+01 | 3.93e-71 | δ1.1 |
| GO:0042254 ribosome biogenesis | 3.98e+01 | 7.31e-67 | δ1.1 |
| GO:0090150 establishment of protein localization to membrane | 3.62e+01 | 8.11e-65 | δ1.1 |
| GO:0034655 nucleobase-containing compound catabolic process | 3.51e+01 | 4.02e-64 | δ1.1 |
| GO:0019058 viral life cycle | 2.99e+01 | 1.89e-62 | δ1.1 |
| GO:0034470 ncRNA processing | 3.24e+01 | 1.98e-62 | δ1.1 |

| | | | |
|---|---|---|---|
| GO:0046700 heterocycle catabolic process | 3.24e+01 | 1.98e-62 | δ1.1 |
| GO:0044270 cellular nitrogen compound catabolic process | 3.19e+01 | 4.56e-62 | δ1.1 |
| GO:0019439 aromatic compound catabolic process | 3.15e+01 | 8.22e-62 | δ1.1 |
| GO:1901361 organic cyclic compound catabolic process | 2.99e+01 | 1e-60 | δ1.1 |
| GO:0022613 ribonucleoprotein complex biogenesis | 2.78e+01 | 4.15e-59 | δ1.1 |
| GO:0072657 protein localization to membrane | 2.72e+01 | 1.03e-58 | δ1.1 |
| GO:0006412 translation | 2.16e+01 | 4.16e-57 | δ1.1 |
| GO:0043043 peptide biosynthetic process | 2.08e+01 | 3.09e-56 | δ1.1 |
| GO:0034660 ncRNA metabolic process | 2.34e+01 | 1.83e-55 | δ1.1 |
| GO:0043604 amide biosynthetic process | 1.89e+01 | 4.38e-54 | δ1.1 |
| GO:0072594 establishment of protein localization to organelle | 2e+01 | 3.41e-52 | δ1.1 |
| GO:0016071 mRNA metabolic process | 1.98e+01 | 5.25e-52 | δ1.1 |
| GO:0006518 peptide metabolic process | 1.7e+01 | 8.71e-52 | δ1.1 |
| GO:1902582 single-organism intracellular transport | 1.92e+01 | 2.65e-51 | δ1.1 |
| GO:0006605 protein targeting | 1.9e+01 | 4.01e-51 | δ1.1 |
| GO:0044802 single-organism membrane organization | 1.57e+01 | 1.29e-48 | δ1.1 |
| GO:0016032 viral process | 1.44e+01 | 3.45e-48 | δ1.1 |
| GO:0044764 multi-organism cellular process | 1.43e+01 | 4.93e-48 | δ1.1 |
| GO:0043603 cellular amide metabolic process | 1.4e+01 | 1.42e-47 | δ1.1 |
| GO:0044419 interspecies interaction between organisms | 1.39e+01 | 1.74e-47 | δ1.1 |
| GO:0044403 symbiosis, encompassing mutualism through parasitism | 1.39e+01 | 1.74e-47 | δ1.1 |
| GO:0033365 protein localization to organelle | 1.48e+01 | 6.9e-46 | δ1.1 |
| GO:0006396 RNA processing | 1.46e+01 | 1.05e-45 | δ1.1 |
| GO:0061024 membrane organization | 1.32e+01 | 5.1e-45 | δ1.1 |

| | | | |
|---|---|---|---|
| GO:0044265 cellular macromolecule catabolic process | 1.36e+01 | 2.91e-44 | δ1.1 |
| GO:0006886 intracellular protein transport | 1.31e+01 | 2.35e-43 | δ1.1 |
| GO:1902580 single-organism cellular localization | 1.21e+01 | 9.06e-42 | δ1.1 |
| GO:1901566 organonitrogen compound biosynthetic process | 1.01e+01 | 1.25e-40 | δ1.1 |
| GO:0009057 macromolecule catabolic process | 1.11e+01 | 5.2e-40 | δ1.1 |
| GO:0046907 intracellular transport | 8.36e+00 | 2.88e-34 | δ1.1 |
| GO:0034613 cellular protein localization | 8.33e+00 | 3.42e-34 | δ1.1 |
| GO:0070727 cellular macromolecule localization | 8.26e+00 | 4.99e-34 | δ1.1 |
| GO:0044248 cellular catabolic process | 8.24e+00 | 5.43e-34 | δ1.1 |
| GO:0015031 protein transport | 7.31e+00 | 1.11e-32 | δ1.1 |
| GO:0051649 establishment of localization in cell | 6.79e+00 | 3.74e-31 | δ1.1 |
| GO:0045184 establishment of protein localization | 6.73e+00 | 5.56e-31 | δ1.1 |
| GO:0008104 protein localization | 5.55e+00 | 4.39e-27 | δ1.1 |
| GO:0051641 cellular localization | 5.35e+00 | 2.49e-26 | δ1.1 |
| GO:0033036 macromolecule localization | 4.83e+00 | 2.95e-24 | δ1.1 |
| GO:0032774 RNA biosynthetic process | 3.55e+00 | 5.86e-19 | δ1.1 |
| GO:0034654 nucleobase-containing compound biosynthetic process | 3.18e+00 | 9.02e-17 | δ1.1 |
| GO:0018130 heterocycle biosynthetic process | 3.15e+00 | 1.58e-16 | δ1.1 |
| GO:0019438 aromatic compound biosynthetic process | 3.14e+00 | 1.81e-16 | δ1.1 |
| GO:0034645 cellular macromolecule biosynthetic process | 2.88e+00 | 3.32e-16 | δ1.1 |
| GO:0016070 RNA metabolic process | 2.99e+00 | 1.72e-15 | δ1.1 |
| GO:0010467 gene expression | 2.72e+00 | 5.02e-15 | δ1.1 |
| GO:0042255 ribosome assembly | 5.09e+01 | 1.1e-11 | δ1.1 |
| GO:0042273 ribosomal large subunit biogenesis | 4.56e+01 | 3.56e-11 | δ1.1 |
| GO:0002181 cytoplasmic translation | 4.71e+01 | 1.42e-08 | δ1.1 |

| | | | |
|---|---|---|---|
| GO:0042274 ribosomal small subunit biogenesis | 3.67e+01 | 1.14e-07 | δ1.1 |
| GO:0022618 ribonucleoprotein complex assembly | 1.57e+01 | 2.16e-07 | δ1.1 |
| GO:0071826 ribonucleoprotein complex subunit organization | 1.49e+01 | 3.71e-07 | δ1.1 |
| GO:0000028 ribosomal small subunit assembly | 7.93e+01 | 1.1e-05 | δ1.1 |
| GO:0000027 ribosomal large subunit assembly | 6.66e+01 | 2.83e-05 | δ1.1 |
| GO:0034622 cellular macromolecular complex assembly | 4.15e+00 | 9.48e-03 | δ1.1 |
| GO:0070925 organelle assembly | 5.1e+00 | 1.85e-02 | δ1.1 |
| GO:0030490 maturation of SSU-rRNA | 2.89e+01 | 2.91e-02 | δ1.1 |
| GO:0000462 maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) | 3e+01 | 3.03e-01 | δ1.1 |
| GO:0000470 maturation of LSU-rRNA | 3.33e+01 | 9.85e-01 | δ1.1 |
| GO:0097421 liver regeneration | 2.87e+01 | 9.97e-01 | δ1.1 |
| GO:0006955 immune response | 3.81e+00 | 3.61e-06 | δ1.2 |
| GO:0002764 immune response-regulating signaling pathway | 6.19e+00 | 2.24e-04 | δ1.2 |
| GO:0002768 immune response-regulating cell surface receptor signaling pathway | 7.14e+00 | 5.11e-04 | δ1.2 |
| hsa04810:Regulation of actin cytoskeleton | 8.58e+00 | 3.09e-05 | δ1.2 |
| GO:0098602 single organism cell adhesion | 4.78e+00 | 7.29e-04 | δ1.2 |
| GO:0001775 cell activation | 4.33e+00 | 1.11e-03 | δ1.2 |
| GO:0050776 regulation of immune response | 4.32e+00 | 1.14e-03 | δ1.2 |
| GO:0016337 single organismal cell-cell adhesion | 4.83e+00 | 1.66e-03 | δ1.2 |
| GO:0038093 Fc receptor signaling pathway | 9.02e+00 | 3.53e-03 | δ1.2 |
| GO:0007159 leukocyte cell-cell adhesion | 5.8e+00 | 4.58e-03 | δ1.2 |
| GO:0002682 regulation of immune system process | 3.34e+00 | 4.79e-03 | δ1.2 |
| GO:0002757 immune response-activating signal transduction | 5.72e+00 | 5.3e-03 | δ1.2 |
| GO:0042110 T cell activation | 5.89e+00 | 1.18e-02 | δ1.2 |

| | | | |
|---|---|---|---|
| GO:0070489 T cell aggregation | 5.89e+00 | 1.18e-02 | δ1.2 |
| GO:0071593 lymphocyte aggregation | 5.88e+00 | 1.2e-02 | δ1.2 |
| GO:0046649 lymphocyte activation | 4.77e+00 | 1.31e-02 | δ1.2 |
| GO:0002429 immune response-activating cell surface receptor signaling pathway | 6.57e+00 | 1.39e-02 | δ1.2 |
| GO:0070486 leukocyte aggregation | 5.79e+00 | 1.39e-02 | δ1.2 |
| GO:0002253 activation of immune response | 5.18e+00 | 1.46e-02 | δ1.2 |
| hsa04666:Fc gamma R-mediated phagocytosis | 1.36e+01 | 9.81e-04 | δ1.2 |
| GO:0098609 cell-cell adhesion | 3.39e+00 | 3.08e-02 | δ1.2 |
| GO:0007015 actin filament organization | 6.54e+00 | 4.65e-02 | δ1.2 |
| GO:0002684 positive regulation of immune system process | 3.65e+00 | 4.98e-02 | δ1.2 |
| GO:0045321 leukocyte activation | 4.11e+00 | 6.31e-02 | δ1.2 |
| GO:0038094 Fc-gamma receptor signaling pathway | 1.17e+01 | 6.62e-02 | δ1.2 |
| GO:0006928 movement of cell or subcellular component | 2.69e+00 | 6.99e-02 | δ1.2 |
| GO:0008219 cell death | 2.57e+00 | 7.9e-02 | δ1.2 |
| GO:0006909 phagocytosis | 7.12e+00 | 8.21e-02 | δ1.2 |
| hsa05131:Shigellosis | 1.54e+01 | 3.82e-03 | δ1.2 |
| GO:0030036 actin cytoskeleton organization | 4.75e+00 | 8.43e-02 | δ1.2 |
| GO:0030217 T cell differentiation | 8.59e+00 | 8.8e-02 | δ1.2 |
| GO:0030029 actin filament-based process | 4.24e+00 | 1.02e-01 | δ1.2 |
| GO:0007155 cell adhesion | 2.71e+00 | 1.03e-01 | δ1.2 |
| GO:0022610 biological adhesion | 2.7e+00 | 1.08e-01 | δ1.2 |
| GO:0050778 positive regulation of immune response | 4.16e+00 | 1.2e-01 | δ1.2 |
| hsa04520:Adherens junction | 1.38e+01 | 6.31e-03 | δ1.2 |
| GO:0007166 cell surface receptor signaling pathway | 2.22e+00 | 1.39e-01 | δ1.2 |
| GO:0006915 apoptotic process | 2.63e+00 | 1.49e-01 | δ1.2 |
| GO:0002252 immune effector process | 3.84e+00 | 2.43e-01 | δ1.2 |
| GO:0071822 protein complex subunit organization | 2.62e+00 | 2.53e-01 | δ1.2 |
| hsa05132:Salmonella infection | 1.18e+01 | 1.33e-02 | δ1.2 |

| | | | |
|---|---|---|---|
| GO:0043933 macromolecular complex subunit organization | 2.23e+00 | 2.75e-01 | δ1.2 |
| GO:0012501 programmed cell death | 2.48e+00 | 3.06e-01 | δ1.2 |
| GO:0008154 actin polymerization or depolymerization | 8.11e+00 | 4.07e-01 | δ1.2 |
| hsa05130:Pathogenic Escherichia coli infection | 1.61e+01 | 2.43e-02 | δ1.2 |
| GO:0038096 Fc-gamma receptor signaling pathway involved in phagocytosis | 1.04e+01 | 4.91e-01 | δ1.2 |
| GO:0002433 immune response-regulating cell surface receptor signaling pathway involved in phagocytosis | 1.04e+01 | 4.91e-01 | δ1.2 |
| GO:0002521 leukocyte differentiation | 4.62e+00 | 4.92e-01 | δ1.2 |
| GO:0002431 Fc receptor mediated stimulatory signaling pathway | 1.01e+01 | 5.28e-01 | δ1.2 |
| GO:0010941 regulation of cell death | 2.61e+00 | 5.48e-01 | δ1.2 |
| GO:0042981 regulation of apoptotic process | 2.66e+00 | 6.51e-01 | δ1.2 |
| GO:0042127 regulation of cell proliferation | 2.54e+00 | 6.68e-01 | δ1.2 |
| GO:0043067 regulation of programmed cell death | 2.64e+00 | 6.89e-01 | δ1.2 |
| GO:0030098 lymphocyte differentiation | 5.65e+00 | 7.06e-01 | δ1.2 |
| hsa04670:Leukocyte transendothelial migration | 8.55e+00 | 5.97e-02 | δ1.2 |
| GO:0048013 ephrin receptor signaling pathway | 1.26e+01 | 7.98e-01 | δ1.2 |
| GO:0035567 non-canonical Wnt signaling pathway | 8.48e+00 | 8.14e-01 | δ1.2 |
| GO:0016477 cell migration | 2.78e+00 | 8.4e-01 | δ1.2 |
| GO:0016192 vesicle-mediated transport | 2.49e+00 | 8.83e-01 | δ1.2 |
| GO:0060491 regulation of cell projection assembly | 8.02e+00 | 8.86e-01 | δ1.2 |
| GO:0044087 regulation of cellular component biogenesis | 3.26e+00 | 9.03e-01 | δ1.2 |
| GO:0008064 regulation of actin polymerization or depolymerization | 7.73e+00 | 9.22e-01 | δ1.2 |
| GO:0030832 regulation of actin filament length | 7.69e+00 | 9.27e-01 | δ1.2 |

| | | | |
|---|---|---|---|
| GO:0051128 regulation of cellular component organization | 2.07e+00 | 9.34e-01 | δ1.2 |
| hsa05100:Bacterial invasion of epithelial cells | 1.05e+01 | 1.18e-01 | δ1.2 |
| GO:0008283 cell proliferation | 2.24e+00 | 9.58e-01 | δ1.2 |
| GO:0048584 positive regulation of response to stimulus | 2.16e+00 | 9.65e-01 | δ1.2 |
| hsa04015:Rap1 signaling pathway | 5.46e+00 | 1.43e-01 | δ1.2 |
| GO:0021762 substantia nigra development | 1.79e+01 | 9.7e-01 | δ1.2 |
| GO:0008284 positive regulation of cell proliferation | 3.1e+00 | 9.7e-01 | δ1.2 |
| GO:0031328 positive regulation of cellular biosynthetic process | 2.28e+00 | 9.76e-01 | δ1.2 |
| GO:0043254 regulation of protein complex assembly | 4.6e+00 | 9.83e-01 | δ1.2 |
| GO:0030097 hemopoiesis | 3.26e+00 | 9.86e-01 | δ1.2 |
| GO:0006461 protein complex assembly | 2.43e+00 | 9.87e-01 | δ1.2 |
| GO:0022407 regulation of cell-cell adhesion | 4.54e+00 | 9.88e-01 | δ1.2 |
| GO:0070271 protein complex biogenesis | 2.42e+00 | 9.88e-01 | δ1.2 |
| GO:0071103 DNA conformation change | 5.4e+00 | 9.88e-01 | δ1.2 |
| GO:0033043 regulation of organelle organization | 2.66e+00 | 9.88e-01 | δ1.2 |
| GO:0009891 positive regulation of biosynthetic process | 2.24e+00 | 9.89e-01 | δ1.2 |
| GO:0001649 osteoblast differentiation | 6.74e+00 | 9.91e-01 | δ1.2 |
| GO:0022607 cellular component assembly | 1.92e+00 | 9.95e-01 | δ1.2 |
| GO:0048870 cell motility | 2.47e+00 | 9.96e-01 | δ1.2 |
| GO:0051674 localization of cell | 2.47e+00 | 9.96e-01 | δ1.2 |
| GO:0051881 regulation of mitochondrial membrane potential | 1.51e+01 | 9.97e-01 | δ1.2 |
| GO:0050863 regulation of T cell activation | 5.13e+00 | 9.97e-01 | δ1.2 |
| GO:0030155 regulation of cell adhesion | 3.41e+00 | 9.97e-01 | δ1.2 |
| GO:0048534 hematopoietic or lymphoid organ development | 3.1e+00 | 9.98e-01 | δ1.2 |
| GO:0051249 regulation of lymphocyte activation | 4.26e+00 | 9.98e-01 | δ1.2 |
| GO:0033151 V(D)J recombination | 3.87e+01 | 9.99e-01 | δ1.2 |

| | | | |
|---|---|---|---|
| GO:0006977 DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest | 1.41e+01 | 9.99e-01 | δ1.2 |
| GO:0006952 defense response | 2.31e+00 | 9.99e-01 | δ1.2 |
| GO:1903037 regulation of leukocyte cell-cell adhesion | 4.9e+00 | 9.99e-01 | δ1.2 |
| GO:1902400 intracellular signal transduction involved in G1 DNA damage checkpoint | 1.39e+01 | 9.99e-01 | δ1.2 |
| GO:0072413 signal transduction involved in mitotic cell cycle checkpoint | 1.39e+01 | 9.99e-01 | δ1.2 |
| GO:1902402 signal transduction involved in mitotic DNA damage checkpoint | 1.39e+01 | 9.99e-01 | δ1.2 |
| GO:1902403 signal transduction involved in mitotic DNA integrity checkpoint | 1.39e+01 | 9.99e-01 | δ1.2 |
| GO:0072431 signal transduction involved in mitotic G1 DNA damage checkpoint | 1.39e+01 | 9.99e-01 | δ1.2 |
| GO:0006897 endocytosis | 3.29e+00 | 9.99e-01 | δ1.2 |
| GO:0010628 positive regulation of gene expression | 2.2e+00 | 1e+00 | δ1.2 |
| GO:0072401 signal transduction involved in DNA integrity checkpoint | 1.35e+01 | 1e+00 | δ1.2 |
| GO:0072422 signal transduction involved in DNA damage checkpoint | 1.35e+01 | 1e+00 | δ1.2 |
| GO:0072395 signal transduction involved in cell cycle checkpoint | 1.33e+01 | 1e+00 | δ1.2 |
| GO:0050851 antigen receptor-mediated signaling pathway | 5.77e+00 | 1e+00 | δ1.2 |
| GO:2000045 regulation of G1/S transition of mitotic cell cycle | 7.77e+00 | 1e+00 | δ1.2 |
| hsa05416:Viral myocarditis | 1.15e+01 | 4.01e-01 | δ1.2 |
| hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC) | 9.78e+00 | 5.53e-01 | δ1.2 |

**Table A.8:** List of GO Biological Process and KEGG pathway terms differentially enriched between δ2.2 and δ2.3 γδ-T cell subtypes in PBMC.

| Term | Fold Enrichment | Bonferroni | cluster |
|---|---|---|---|
| GO:0019835 cytolysis | 7.67e+01 | 4.69e-01 | δ2.2 |
| GO:0002703 regulation of leukocyte mediated immunity | 2.03e+01 | 5.64e-01 | δ2.2 |
| GO:0002699 positive regulation of immune effector process | 1.93e+01 | 6.17e-01 | δ2.2 |
| hsa04612:Antigen processing and presentation | 2.47e+01 | 8.35e-02 | δ2.2 |
| GO:0006614 SRP-dependent cotranslational protein targeting to membrane | 6.35e+01 | 4.51e-19 | δ2.3 |
| GO:0006613 cotranslational protein targeting to membrane | 5.91e+01 | 1.28e-18 | δ2.3 |
| GO:0045047 protein targeting to ER | 5.86e+01 | 1.48e-18 | δ2.3 |
| GO:0072599 establishment of protein localization to endoplasmic reticulum | 5.64e+01 | 2.58e-18 | δ2.3 |
| GO:0000184 nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 4.95e+01 | 1.74e-17 | δ2.3 |
| GO:0070972 protein localization to endoplasmic reticulum | 4.76e+01 | 3.12e-17 | δ2.3 |
| GO:0019083 viral transcription | 3.37e+01 | 4.57e-15 | δ2.3 |
| GO:0006413 translational initiation | 3.22e+01 | 8.46e-15 | δ2.3 |
| GO:0006612 protein targeting to membrane | 3.21e+01 | 9.12e-15 | δ2.3 |
| GO:0019080 viral gene expression | 3.17e+01 | 1.06e-14 | δ2.3 |
| GO:0000956 nuclear-transcribed mRNA catabolic process | 2.96e+01 | 2.87e-14 | δ2.3 |
| GO:0044033 multi-organism metabolic process | 2.85e+01 | 4.91e-14 | δ2.3 |
| hsa03010:Ribosome | 2.53e+01 | 1.51e-15 | δ2.3 |
| GO:0006402 mRNA catabolic process | 2.76e+01 | 1.47e-13 | δ2.3 |
| GO:0006401 RNA catabolic process | 2.45e+01 | 4.42e-13 | δ2.3 |
| GO:0006364 rRNA processing | 2.26e+01 | 1.18e-12 | δ2.3 |
| GO:0016072 rRNA metabolic process | 2.21e+01 | 1.77e-12 | δ2.3 |

| | | | |
|---|---|---|---|
| GO:0090150 establishment of protein localization to membrane | 1.77e+01 | 3.53e-12 | δ2.3 |
| GO:0042254 ribosome biogenesis | 1.82e+01 | 2.61e-11 | δ2.3 |
| GO:1901361 organic cyclic compound catabolic process | 1.46e+01 | 5.8e-11 | δ2.3 |
| GO:0034655 nucleobase-containing compound catabolic process | 1.6e+01 | 1.47e-10 | δ2.3 |
| GO:0072657 protein localization to membrane | 1.33e+01 | 2.29e-10 | δ2.3 |
| GO:0046700 heterocycle catabolic process | 1.48e+01 | 4.3e-10 | δ2.3 |
| GO:0034470 ncRNA processing | 1.48e+01 | 4.3e-10 | δ2.3 |
| GO:0044270 cellular nitrogen compound catabolic process | 1.46e+01 | 5.41e-10 | δ2.3 |
| GO:0019439 aromatic compound catabolic process | 1.44e+01 | 6.36e-10 | δ2.3 |
| GO:0019058 viral life cycle | 1.34e+01 | 1.71e-09 | δ2.3 |
| GO:0022613 ribonucleoprotein complex biogenesis | 1.27e+01 | 3.52e-09 | δ2.3 |
| GO:0006412 translation | 1.01e+01 | 1.12e-08 | δ2.3 |
| GO:0072594 establishment of protein localization to organelle | 9.76e+00 | 1.93e-08 | δ2.3 |
| GO:0043043 peptide biosynthetic process | 9.74e+00 | 1.98e-08 | δ2.3 |
| GO:0016071 mRNA metabolic process | 9.67e+00 | 2.2e-08 | δ2.3 |
| GO:0006518 peptide metabolic process | 8.44e+00 | 2.74e-08 | δ2.3 |
| GO:0034660 ncRNA metabolic process | 1.07e+01 | 3.51e-08 | δ2.3 |
| GO:0043604 amide biosynthetic process | 8.83e+00 | 7.94e-08 | δ2.3 |
| GO:0044265 cellular macromolecule catabolic process | 7.07e+00 | 3.78e-07 | δ2.3 |
| GO:1902582 single-organism intracellular transport | 8.76e+00 | 4.81e-07 | δ2.3 |
| GO:0043603 cellular amide metabolic process | 6.95e+00 | 4.95e-07 | δ2.3 |
| GO:0006605 protein targeting | 8.69e+00 | 5.38e-07 | δ2.3 |
| GO:0006886 intracellular protein transport | 6.77e+00 | 7.23e-07 | δ2.3 |
| GO:0044802 single-organism membrane organization | 7.5e+00 | 7.69e-07 | δ2.3 |

| | | | |
|---|---|---|---|
| GO:0033365 protein localization to organelle | 7.2e+00 | 1.35e-06 | δ2.3 |
| GO:0006396 RNA processing | 7.14e+00 | 1.52e-06 | δ2.3 |
| GO:0016032 viral process | 6.73e+00 | 3.46e-06 | δ2.3 |
| GO:0044764 multi-organism cellular process | 6.68e+00 | 3.82e-06 | δ2.3 |
| GO:0044419 interspecies interaction between organisms | 6.51e+00 | 5.4e-06 | δ2.3 |
| GO:0044403 symbiosis, encompassing mutualism through parasitism | 6.51e+00 | 5.4e-06 | δ2.3 |
| GO:0009057 macromolecule catabolic process | 5.75e+00 | 7.74e-06 | δ2.3 |
| GO:0061024 membrane organization | 6.33e+00 | 7.97e-06 | δ2.3 |
| GO:1902580 single-organism cellular localization | 5.9e+00 | 2.07e-05 | δ2.3 |
| GO:1901566 organonitrogen compound biosynthetic process | 5.02e+00 | 5.31e-05 | δ2.3 |
| GO:0046907 intracellular transport | 4.33e+00 | 4.14e-04 | δ2.3 |
| GO:0034613 cellular protein localization | 4.31e+00 | 4.36e-04 | δ2.3 |
| GO:0070727 cellular macromolecule localization | 4.28e+00 | 4.87e-04 | δ2.3 |
| GO:0044248 cellular catabolic process | 4.27e+00 | 5e-04 | δ2.3 |
| GO:0015031 protein transport | 3.71e+00 | 3.32e-03 | δ2.3 |
| GO:0051649 establishment of localization in cell | 3.44e+00 | 8.84e-03 | δ2.3 |
| GO:0045184 establishment of protein localization | 3.42e+00 | 9.86e-03 | δ2.3 |
| GO:0042110 T cell activation | 8.19e+00 | 1.19e-02 | δ2.3 |
| GO:0070489 T cell aggregation | 8.19e+00 | 1.19e-02 | δ2.3 |
| GO:0071593 lymphocyte aggregation | 8.18e+00 | 1.21e-02 | δ2.3 |
| GO:0070486 leukocyte aggregation | 8.05e+00 | 1.35e-02 | δ2.3 |
| GO:0098609 cell-cell adhesion | 4.54e+00 | 1.51e-02 | δ2.3 |
| GO:0042255 ribosome assembly | 3.38e+01 | 1.66e-02 | δ2.3 |
| GO:0046649 lymphocyte activation | 6.32e+00 | 2.21e-02 | δ2.3 |
| GO:0007159 leukocyte cell-cell adhesion | 7.44e+00 | 2.37e-02 | δ2.3 |
| GO:0001913 T cell mediated cytotoxicity | 5.41e+01 | 6.4e-02 | δ2.3 |
| GO:0045321 leukocyte activation | 5.44e+00 | 7.01e-02 | δ2.3 |
| GO:0002250 adaptive immune response | 7.74e+00 | 7.17e-02 | δ2.3 |

| | | | |
|---|---|---|---|
| GO:0008104 protein localization | 2.82e+00 | 1.06e-01 | δ2.3 |
| GO:0030217 T cell differentiation | 1.19e+01 | 1.49e-01 | δ2.3 |
| GO:0051641 cellular localization | 2.72e+00 | 1.62e-01 | δ2.3 |
| GO:0033993 response to lipid | 4.72e+00 | 1.93e-01 | δ2.3 |
| GO:0070887 cellular response to chemical stimulus | 2.61e+00 | 2.51e-01 | δ2.3 |
| GO:0001775 cell activation | 4.46e+00 | 2.79e-01 | δ2.3 |
| GO:0016337 single organismal cell-cell adhesion | 5.04e+00 | 3.08e-01 | δ2.3 |
| GO:0002285 lymphocyte activation involved in immune response | 1.36e+01 | 4.41e-01 | δ2.3 |
| GO:0048534 hematopoietic or lymphoid organ development | 4.7e+00 | 4.46e-01 | δ2.3 |
| GO:0098602 single organism cell adhesion | 4.69e+00 | 4.49e-01 | δ2.3 |
| GO:0007155 cell adhesion | 3.11e+00 | 4.49e-01 | δ2.3 |
| GO:0033036 macromolecule localization | 2.45e+00 | 4.5e-01 | δ2.3 |
| GO:0022610 biological adhesion | 3.1e+00 | 4.6e-01 | δ2.3 |
| GO:0042273 ribosomal large subunit biogenesis | 2.42e+01 | 5.2e-01 | δ2.3 |
| GO:0002460 adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains | 8.52e+00 | 5.37e-01 | δ2.3 |
| GO:0002520 immune system development | 4.45e+00 | 5.7e-01 | δ2.3 |
| GO:0044085 cellular component biogenesis | 2.37e+00 | 5.86e-01 | δ2.3 |
| GO:0010467 gene expression | 1.84e+00 | 6.11e-01 | δ2.3 |
| GO:0006954 inflammatory response | 5.1e+00 | 6.15e-01 | δ2.3 |
| GO:0032774 RNA biosynthetic process | 2.08e+00 | 6.25e-01 | δ2.3 |
| GO:0048535 lymph node development | 7.17e+01 | 6.29e-01 | δ2.3 |
| GO:0030098 lymphocyte differentiation | 7.86e+00 | 6.7e-01 | δ2.3 |
| GO:0002521 leukocyte differentiation | 6e+00 | 6.85e-01 | δ2.3 |
| GO:0002456 T cell mediated immunity | 2.06e+01 | 6.95e-01 | δ2.3 |
| GO:0001909 leukocyte mediated cytotoxicity | 1.98e+01 | 7.34e-01 | δ2.3 |
| GO:0034654 nucleobase-containing compound biosynthetic process | 1.96e+00 | 7.54e-01 | δ2.3 |
| GO:0018130 heterocycle biosynthetic process | 1.93e+00 | 8.09e-01 | δ2.3 |

| | | | |
|---|---|---|---|
| GO:0002286 T cell activation involved in immune response | 1.83e+01 | 8.13e-01 | δ2.3 |
| GO:0019438 aromatic compound biosynthetic process | 1.93e+00 | 8.21e-01 | δ2.3 |
| GO:0001914 regulation of T cell mediated cytotoxicity | 5.3e+01 | 8.4e-01 | δ2.3 |
| GO:0000027 ribosomal large subunit assembly | 4.87e+01 | 8.85e-01 | δ2.3 |
| GO:0002366 leukocyte activation involved in immune response | 9.58e+00 | 8.85e-01 | δ2.3 |
| GO:0022618 ribonucleoprotein complex assembly | 9.58e+00 | 8.85e-01 | δ2.3 |
| GO:0002263 cell activation involved in immune response | 9.49e+00 | 8.94e-01 | δ2.3 |
| GO:0030097 hemopoiesis | 4.4e+00 | 8.96e-01 | δ2.3 |
| GO:0071826 ribonucleoprotein complex subunit organization | 9.11e+00 | 9.26e-01 | δ2.3 |
| GO:0016070 RNA metabolic process | 1.83e+00 | 9.63e-01 | δ2.3 |
| GO:0051249 regulation of lymphocyte activation | 5.93e+00 | 9.78e-01 | δ2.3 |
| GO:0046651 lymphocyte proliferation | 7.72e+00 | 9.91e-01 | δ2.3 |
| GO:0032943 mononuclear cell proliferation | 7.66e+00 | 9.92e-01 | δ2.3 |
| GO:0071310 cellular response to organic substance | 2.42e+00 | 9.96e-01 | δ2.3 |
| GO:0070661 leukocyte proliferation | 7.25e+00 | 9.97e-01 | δ2.3 |
| GO:0010033 response to organic substance | 2.16e+00 | 9.97e-01 | δ2.3 |
| GO:0002694 regulation of leukocyte activation | 5.21e+00 | 9.99e-01 | δ2.3 |
| GO:0050863 regulation of T cell activation | 6.79e+00 | 9.99e-01 | δ2.3 |
| GO:0002293 alpha-beta T cell differentiation involved in immune response | 2.44e+01 | 1e+00 | δ2.3 |
| GO:0002287 alpha-beta T cell activation involved in immune response | 2.44e+01 | 1e+00 | δ2.3 |
| GO:1903037 regulation of leukocyte cell-cell adhesion | 6.49e+00 | 1e+00 | δ2.3 |
| GO:0001910 regulation of leukocyte mediated cytotoxicity | 2.39e+01 | 1e+00 | δ2.3 |

| | | | |
|---|---|---|---|
| GO:0050865 regulation of cell activation | 4.86e+00 | 1e+00 | δ2.3 |
| GO:0002709 regulation of T cell mediated immunity | 2.34e+01 | 1e+00 | δ2.3 |
| GO:1901700 response to oxygen-containing compound | 2.77e+00 | 1e+00 | δ2.3 |
| GO:0002684 positive regulation of immune system process | 3.38e+00 | 1e+00 | δ2.3 |
| GO:0002440 production of molecular mediator of immune response | 9.73e+00 | 1e+00 | δ2.3 |
| GO:0002292 T cell differentiation involved in immune response | 2.18e+01 | 1e+00 | δ2.3 |
| GO:0031341 regulation of cell killing | 2e+01 | 1e+00 | δ2.3 |
| GO:0045595 regulation of cell differentiation | 2.63e+00 | 1e+00 | δ2.3 |

**Table A.9:** List of GO Biological Process and KEGG pathway terms differentially enriched γδ-T cell subtypes in breast tumour.

| Term | Fold Enrichment | Bonferroni | cluster |
|---|---|---|---|
| GO:0001775 cell activation | 3.47e+00 | 4.64e-09 | γδ-T.1 |
| GO:0007166 cell surface receptor signaling pathway | 2.1e+00 | 2.24e-07 | γδ-T.1 |
| GO:0050776 regulation of immune response | 3.07e+00 | 4.31e-06 | γδ-T.1 |
| GO:0002764 immune response-regulating signaling pathway | 3.92e+00 | 5.54e-06 | γδ-T.1 |
| GO:0045321 leukocyte activation | 3.27e+00 | 1.48e-05 | γδ-T.1 |
| GO:0002757 immune response-activating signal transduction | 3.89e+00 | 2.76e-05 | γδ-T.1 |
| GO:0002768 immune response-regulating cell surface receptor signaling pathway | 4.32e+00 | 3.47e-05 | γδ-T.1 |
| GO:0002252 immune effector process | 3.1e+00 | 1.68e-04 | γδ-T.1 |
| GO:0002429 immune response-activating cell surface receptor signaling pathway | 4.3e+00 | 1.81e-04 | γδ-T.1 |
| GO:0048584 positive regulation of response to stimulus | 2.09e+00 | 1.82e-04 | γδ-T.1 |
| GO:0002253 activation of immune response | 3.52e+00 | 2.05e-04 | γδ-T.1 |
| GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway | 3.21e+00 | 4.03e-04 | γδ-T.1 |
| hsa04650:Natural killer cell mediated cytotoxicity | 6.75e+00 | 2.14e-05 | γδ-T.1 |
| GO:0009966 regulation of signal transduction | 1.84e+00 | 8.07e-04 | γδ-T.1 |
| GO:0036211 protein modification process | 1.65e+00 | 1.14e-03 | γδ-T.1 |
| GO:0006464 cellular protein modification process | 1.65e+00 | 1.14e-03 | γδ-T.1 |
| GO:0006909 phagocytosis | 4.66e+00 | 1.37e-03 | γδ-T.1 |
| GO:0007167 enzyme linked receptor protein signaling pathway | 2.63e+00 | 1.56e-03 | γδ-T.1 |
| GO:0016310 phosphorylation | 1.93e+00 | 1.94e-03 | γδ-T.1 |
| GO:0007155 cell adhesion | 2.11e+00 | 2.05e-03 | γδ-T.1 |
| GO:0022610 biological adhesion | 2.1e+00 | 2.28e-03 | γδ-T.1 |

| | | | |
|---|---|---|---|
| GO:0035556 intracellular signal transduction | 1.81e+00 | 3.25e-03 | γδ-T.1 |
| GO:0010646 regulation of cell communication | 1.73e+00 | 3.99e-03 | γδ-T.1 |
| GO:0050778 positive regulation of immune response | 2.94e+00 | 4.08e-03 | γδ-T.1 |
| GO:0046649 lymphocyte activation | 3.01e+00 | 4.21e-03 | γδ-T.1 |
| GO:0008219 cell death | 1.98e+00 | 4.52e-03 | γδ-T.1 |
| GO:0012501 programmed cell death | 2.02e+00 | 4.59e-03 | γδ-T.1 |
| GO:0006915 apoptotic process | 2.05e+00 | 4.64e-03 | γδ-T.1 |
| GO:0023051 regulation of signaling | 1.71e+00 | 7.45e-03 | γδ-T.1 |
| hsa04810:Regulation of actin cytoskeleton | 4.48e+00 | 4.1e-04 | γδ-T.1 |
| GO:0007229 integrin-mediated signaling pathway | 7.45e+00 | 8.73e-03 | γδ-T.1 |
| GO:0006796 phosphate-containing compound metabolic process | 1.7e+00 | 8.73e-03 | γδ-T.1 |
| GO:0006793 phosphorus metabolic process | 1.7e+00 | 9.39e-03 | γδ-T.1 |
| GO:0051674 localization of cell | 2.21e+00 | 1.02e-02 | γδ-T.1 |
| GO:0048870 cell motility | 2.21e+00 | 1.02e-02 | γδ-T.1 |
| GO:0016477 cell migration | 2.3e+00 | 1.07e-02 | γδ-T.1 |
| GO:0018193 peptidyl-amino acid modification | 2.3e+00 | 1.09e-02 | γδ-T.1 |
| hsa04670:Leukocyte transendothelial migration | 6.14e+00 | 6.37e-04 | γδ-T.1 |
| GO:0042127 regulation of cell proliferation | 2.08e+00 | 1.4e-02 | γδ-T.1 |
| GO:0010647 positive regulation of cell communication | 2.08e+00 | 1.47e-02 | γδ-T.1 |
| GO:1902531 regulation of intracellular signal transduction | 2.01e+00 | 1.5e-02 | γδ-T.1 |
| GO:0050851 antigen receptor-mediated signaling pathway | 4.72e+00 | 1.51e-02 | γδ-T.1 |
| GO:2000145 regulation of cell motility | 2.73e+00 | 1.64e-02 | γδ-T.1 |
| GO:0023056 positive regulation of signaling | 2.07e+00 | 1.67e-02 | γδ-T.1 |
| GO:0098609 cell-cell adhesion | 2.28e+00 | 1.88e-02 | γδ-T.1 |
| GO:0009967 positive regulation of signal transduction | 2.11e+00 | 2.35e-02 | γδ-T.1 |
| GO:0002274 myeloid leukocyte activation | 5.86e+00 | 2.61e-02 | γδ-T.1 |

| | | | |
|---|---|---|---|
| GO:0051270 regulation of cellular component movement | 2.6e+00 | 2.68e-02 | γδ-T.1 |
| GO:0038093 Fc receptor signaling pathway | 4.43e+00 | 3.11e-02 | γδ-T.1 |
| GO:0042981 regulation of apoptotic process | 2.1e+00 | 3.48e-02 | γδ-T.1 |
| GO:0040012 regulation of locomotion | 2.62e+00 | 3.49e-02 | γδ-T.1 |
| GO:0002684 positive regulation of immune system process | 2.39e+00 | 3.96e-02 | γδ-T.1 |
| GO:0043067 regulation of programmed cell death | 2.08e+00 | 4.31e-02 | γδ-T.1 |
| GO:0098602 single organism cell adhesion | 2.58e+00 | 4.47e-02 | γδ-T.1 |
| GO:0006468 protein phosphorylation | 1.9e+00 | 5.24e-02 | γδ-T.1 |
| GO:0040011 locomotion | 2.02e+00 | 5.26e-02 | γδ-T.1 |
| GO:0007159 leukocyte cell-cell adhesion | 3.07e+00 | 6.88e-02 | γδ-T.1 |
| GO:0006897 endocytosis | 2.7e+00 | 7.3e-02 | γδ-T.1 |
| GO:0010941 regulation of cell death | 2e+00 | 8.49e-02 | γδ-T.1 |
| hsa04510:Focal adhesion | 4e+00 | 7.76e-03 | γδ-T.1 |
| hsa04015:Rap1 signaling pathway | 3.92e+00 | 9.48e-03 | γδ-T.1 |
| hsa05416:Viral myocarditis | 7.22e+00 | 7.05e-02 | γδ-T.1 |
| hsa04650:Natural killer cell mediated cytotoxicity | 1.21e+01 | 3.33e-07 | γδ-T.2 |
| GO:0051674 localization of cell | 3.23e+00 | 1.03e-05 | γδ-T.2 |
| GO:0048870 cell motility | 3.23e+00 | 1.03e-05 | γδ-T.2 |
| GO:0040011 locomotion | 2.98e+00 | 2.21e-05 | γδ-T.2 |
| GO:0046649 lymphocyte activation | 4.45e+00 | 5.12e-05 | γδ-T.2 |
| GO:0007166 cell surface receptor signaling pathway | 2.29e+00 | 7.04e-05 | γδ-T.2 |
| GO:0006928 movement of cell or subcellular component | 2.68e+00 | 1.03e-04 | γδ-T.2 |
| GO:0045321 leukocyte activation | 4.01e+00 | 1.42e-04 | γδ-T.2 |
| GO:0001775 cell activation | 3.43e+00 | 1.06e-03 | γδ-T.2 |
| GO:0050776 regulation of immune response | 3.28e+00 | 4.3e-03 | γδ-T.2 |
| GO:0016477 cell migration | 2.86e+00 | 8.34e-03 | γδ-T.2 |
| GO:0002252 immune effector process | 3.51e+00 | 1.04e-02 | γδ-T.2 |
| GO:0030334 regulation of cell migration | 3.61e+00 | 1.35e-02 | γδ-T.2 |
| GO:0002228 natural killer cell mediated immunity | 1.52e+01 | 1.89e-02 | γδ-T.2 |

| | | | |
|---|---|---|---|
| GO:0007155 cell adhesion | 2.37e+00 | 2.66e-02 | γδ-T.2 |
| GO:0022610 biological adhesion | 2.37e+00 | 2.85e-02 | γδ-T.2 |
| GO:2000145 regulation of cell motility | 3.36e+00 | 3.57e-02 | γδ-T.2 |
| GO:0040012 regulation of locomotion | 3.22e+00 | 6.24e-02 | γδ-T.2 |
| GO:0031589 cell-substrate adhesion | 5.1e+00 | 6.68e-02 | γδ-T.2 |
| GO:0007159 leukocyte cell-cell adhesion | 3.97e+00 | 6.77e-02 | γδ-T.2 |
| hsa05100:Bacterial invasion of epithelial cells | 9.45e+00 | 5.54e-02 | γδ-T.2 |
| GO:0080134 regulation of response to stress | 3.99e+00 | 2.54e-05 | γδ-T.3 |
| GO:1901700 response to oxygen-containing compound | 3.2e+00 | 5.48e-03 | γδ-T.3 |
| hsa05321:Inflammatory bowel disease (IBD) | 1.75e+01 | 2.87e-04 | γδ-T.3 |
| GO:0009966 regulation of signal transduction | 2.35e+00 | 1.56e-02 | γδ-T.3 |
| GO:0033993 response to lipid | 3.97e+00 | 2.16e-02 | γδ-T.3 |
| GO:0048585 negative regulation of response to stimulus | 3.12e+00 | 3e-02 | γδ-T.3 |
| GO:0010646 regulation of cell communication | 2.19e+00 | 3.92e-02 | γδ-T.3 |
| GO:0023051 regulation of signaling | 2.16e+00 | 5.41e-02 | γδ-T.3 |
| GO:1902531 regulation of intracellular signal transduction | 2.74e+00 | 6.11e-02 | γδ-T.3 |
| GO:0002684 positive regulation of immune system process | 3.56e+00 | 7.7e-02 | γδ-T.3 |
| GO:0048584 positive regulation of response to stimulus | 2.52e+00 | 7.73e-02 | γδ-T.3 |
| GO:0010033 response to organic substance | 2.19e+00 | 8.87e-02 | γδ-T.3 |
| hsa05142:Chagas disease (American trypanosomiasis) | 9.23e+00 | 5.29e-02 | γδ-T.3 |